

## PREDICTING DEVICE PERFORMING PREDICTION BASED ON ANALOGOUS EXAMPLE AND METHOD

Patent number: JP2000155681

Publication date: 2000-06-06

Inventor: MAEDA KAZUO; YAGINUMA YOSHINORI

Applicant: FUJITSU LTD

Classification:

- International: G06N5/00; G06N5/00; (IPC1-7): G06F9/44; G06F9/44

- european: G06N5/00

Application number: JP19980332503 19981124

Priority number(s): JP19980332503 19981124

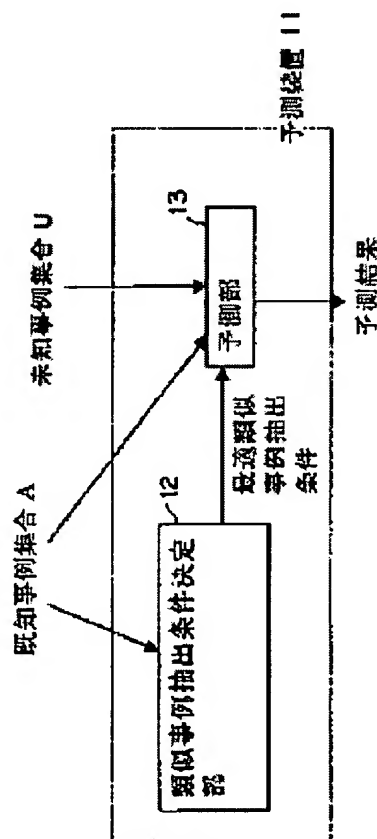
Also published as:

US 6418425 (B1)

Report a data error here

### Abstract of JP2000155681

**PROBLEM TO BE SOLVED:** To perform prediction based on an analogous example fast and also with high accuracy. **SOLUTION:** An analogous example extraction condition deciding part 12 automatically decides an optimum analogous example extraction condition by using a known example set A, and a predicting part 13 predicts an unknown field of an unknown example set U from the known example set A by using the condition. In such a case, the predicting part 13 calculates the similarity depending on the distribution of unknown field values in the set A and extracts an analogous example set based on the similarity. Also, the predicting part 13 stops a similarity calculation at the point of time when it is determined that the similarity does not meet a prescribed condition.



Data supplied from the esp@cenet database - Worldwide

h)

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号  
特開2000-155681  
(P2000-155681A)

(43)公開日 平成12年6月6日(2000.6.6)

(51)Int.Cl.<sup>7</sup>  
G 0 6 F 9/44

識別記号  
5 5 0  
5 8 0

F I  
G 0 6 F 9/44

テームコード(参考)  
5 5 0 N  
5 8 0 J

審査請求 未請求 請求項の数33 O L (全 26 頁)

|          |                         |         |   |
|----------|-------------------------|---------|---|
| (21)出願番号 | 特願平10-332503            | (71)出願人 | 000005223<br>富士通株式会社<br>神奈川県川崎市中原区上小田中4丁目1番1号 |
| (22)出願日  | 平成10年11月24日(1998.11.24) | (72)発明者 | 前田 一穂<br>神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内       |
|          |                         | (72)発明者 | 柳沼 義典<br>神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内       |
|          |                         | (74)代理人 | 100074099<br>弁理士 大菅 義之 (外1名)                  |

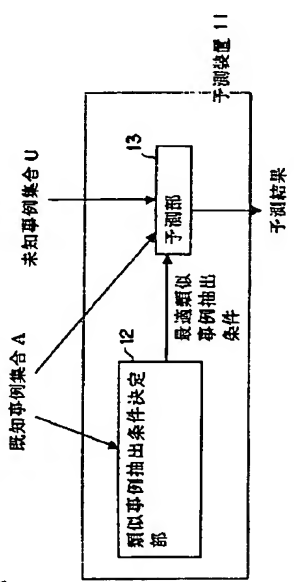
(54)【発明の名称】 類似事例に基づく予測を行う予測装置および方法

(57)【要約】

【課題】 類似事例に基づく予測を高速かつ高精度に行うことが課題である。

【解決手段】 類似事例抽出条件決定部12は、既知事例集合Aを用いて最適な類似事例抽出条件を自動的に決定し、予測部13は、その条件を用いて、既知事例集合Aから未知事例集合Uの未知フィールドを予測する。このとき、予測部13は、既知事例集合Aにおける未知フィールドの値の分布に依存する類似度を計算し、類似度に基づいて類似事例集合を抽出する。また、予測部13は、類似度が所定の条件を満たさないことが確定した時点で、類似度計算を中止する。

予測装置の構成図



## 【特許請求の範囲】

【請求項1】 類似事例に基づく予測を行う予測装置であって、

1つ以上のフィールドからなる既知事例データの集合から未知事例データに類似する1つ以上の類似事例データを抽出するための類似事例抽出条件を自動的に決定する決定手段と、

前記決定手段により決定された類似事例抽出条件を用いて、前記1つ以上の類似事例データを抽出し、該1つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測し、予測値を出力する予測手段とを備えることを特徴とする予測装置。

【請求項2】 前記決定手段は、複数の類似事例抽出条件を評価し、評価結果に基づいて、適切な類似事例抽出条件を選択することを特徴とする請求項1記載の予測装置。

【請求項3】 前記決定手段は、類似事例抽出条件の種類を指定する種類情報と指定された類似事例抽出条件の範囲を指定する範囲情報を出力する条件出力手段と、

前記条件出力手段の出力を包含するような最大条件を求める最大条件計算手段と、

前記既知事例データの集合から既知事例入力用事例データの集合と未知事例入力用事例データの集合を生成する入力用事例生成手段と、

前記最大条件に従って、前記既知事例入力用事例データの集合から前記未知事例入力用事例データに類似する類似事例データの集合を抽出する類似事例抽出手段と、

前記未知事例入力用事例データに類似する類似事例データの集合を用いて、前記類似事例抽出条件の範囲の中から適切な類似事例抽出条件を選択し、該適切な類似事例抽出条件の情報を出力する条件決定手段とを含むことを特徴とする請求項1記載の予測装置。

【請求項4】 前記条件出力手段は、類似事例数を前記種類情報として出力し、該類似事例数の範囲を前記範囲情報として出力することを特徴とする請求項3記載の予測装置。

【請求項5】 前記条件出力手段は、類似度の閾値を前記種類情報として出力し、該閾値の範囲を前記範囲情報として出力することを特徴とする請求項3記載の予測装置。

【請求項6】 前記条件出力手段は、類似事例数および類似度を含む条件式を前記種類情報として出力し、該条件式の検査範囲を前記範囲情報として出力することを特徴とする請求項3記載の予測装置。

【請求項7】 前記入力用事例生成手段は、前記既知事例データの集合を2つに分割して、一方を前記既知事例入力用事例データの集合として出力し、他方を前記未知事例入力用事例データの集合として出力することを特徴とする請求項3記載の予測装置。

【請求項8】 前記入力用事例生成手段は、前記既知事例データの集合を前記未知事例入力用事例データの集合として出力し、該既知事例データの集合から1つの未知事例入力用事例データを削除して得られた集合を、該1つの未知事例入力用事例データに対する既知事例入力用事例データの集合として出力することを特徴とする請求項3記載の予測装置。

【請求項9】 前記入力用事例生成手段は、前記既知事例データの集合から1つ以上の事例データをサンプリングして得られた集合を、前記未知事例入力用事例データの集合として出力し、該既知事例データの集合から1つの未知事例入力用事例データを削除して得られた集合を、該1つの未知事例入力用事例データに対する既知事例入力用事例データの集合として出力することを特徴とする請求項3記載の予測装置。

【請求項10】 前記決定手段は、前記既知事例入力用事例データの集合に前記未知事例入力用事例データと重複する事例データが含まれているとき、前記最大条件を修正して前記類似事例抽出手段に出力する最大条件修正手段と、

前記最大条件が修正されたとき、前記類似事例抽出手段から出力された前記類似事例データの集合から事例データを削除して、修正された類似事例データの集合を前記条件決定手段に出力する類似事例削除手段とをさらに含むことを特徴とする請求項3記載の予測装置。

【請求項11】 前記入力用事例生成手段は、前記既知事例データの集合を前記未知事例入力用事例データの集合および既知事例入力用事例データの集合として出力し、前記最大条件修正手段は、類似事例数が1つ多くなるように前記最大条件を修正し、前記類似事例削除手段は、1つの未知事例入力用事例データに対する類似事例データの集合から該1つの未知事例入力用事例データと重複する事例データを削除することを特徴とする請求項10記載の予測装置。

【請求項12】 前記入力用事例生成手段は、前記既知事例データの集合から1つ以上の事例データをサンプリングして得られた集合を、前記未知事例入力用事例データの集合として出力し、該既知事例データの集合を前記既知事例入力用事例データの集合として出力し、前記最大条件修正手段は、類似事例数が1つ多くなるように前記最大条件を修正し、前記類似事例削除手段は、1つの未知事例入力用事例データに対する類似事例データの集合から該1つの未知事例入力用事例データと重複する事例データを削除することを特徴とする請求項10記載の予測装置。

【請求項13】 前記類似事例抽出手段は、記憶に基づく推論および事例に基づく推論のうちのいずれかを用いて、前記未知事例入力用事例データに類似する類似事例データの集合を抽出することを特徴とする請求項3記載の予測装置。

【請求項14】 前記条件決定手段は、前記類似事例抽出条件の範囲に含まれる条件を離散化して出力する条件離散化手段と、前記未知事例入力用事例データに類似する類似事例データの集合から、離散化された条件の各々に合致する条件ごとの類似事例データの集合を抽出する条件付き類似事例抽出手段と、前記条件ごとの類似事例データの集合を用いて、条件ごとに前記未知事例入力用事例データの予測対象フィールドの値を予測し、条件ごとの予測値を出力する予測結果生成手段と、前記条件ごとの予測値から条件ごとの評価値を求める条件評価手段と、前記条件ごとの評価値に基づいて、前記離散化された条件の中から前記適切な類似事例抽出条件を選択する条件選択手段とを含むことを特徴とする請求項3記載の予測装置。

【請求項15】 前記予測結果生成手段は、記憶に基づく推論および事例に基づく推論のうちのいずれかを用いて、前記予測結果を生成することを特徴とする請求項14記載の予測装置。

【請求項16】 前記条件評価手段は、前記予測値がカテゴリ値であるとき、該予測値と前記予測対象フィールドの真の値を比較し、該予測値と真の値が一致するかどうかに従って前記評価値を生成することを特徴とする請求項14記載の予測装置。

【請求項17】 前記条件評価手段は、前記予測値が連続値であるとき、該予測値と前記予測対象フィールドの真の値を比較し、該予測値と真の値の差を用いて前記評価値を生成することを特徴とする請求項14記載の予測装置。

【請求項18】 前記条件評価手段は、前記予測値に付随する確信度を加味して前記評価値を生成することを特徴とする請求項14記載の予測装置。

【請求項19】 前記条件評価手段は、前記予測値と前記予測対象フィールドの真の値から重みを求め、該重みを加味して前記評価値を生成することを特徴とする請求項14記載の予測装置。

【請求項20】 前記条件評価手段は、1つの離散化された条件と該1つの離散化された条件に合致する類似事例データの集合の少なくとも一方を用いて、類似事例抽出の実行時間を推定し、推定された実行時間を加味して前記評価値を生成することを特徴とする請求項14記載の予測装置。

【請求項21】 前記条件選択手段は、与えられた評価値のうち最良の値に対応する条件を前記適切な類似事例抽出条件として選択することを特徴とする請求項14記載の予測装置。

【請求項22】 前記条件選択手段は、与えられた評価値について移動平均を計算し、得られた平均評価値のう

ち最良の値に対応する条件を前記適切な類似事例抽出条件として選択することを特徴とする請求項14記載の予測装置。

【請求項23】 前記条件選択手段は、与えられた評価値を条件の関数により近似し、得られた近似評価値のうち最良の値に対応する条件を前記適切な類似事例抽出条件として選択することを特徴とする請求項14記載の予測装置。

【請求項24】 類似事例に基づく予測を行う予測装置であって、

1つ以上のフィールドからなる既知事例データの集合から、類似度に基づいて未知事例データに類似する1つ以上の類似事例データを抽出する類似事例抽出手段と、

前記1つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測し、予測値を出力する予測結果生成手段と、

前記既知事例データの集合における前記未知フィールドの値の分布と前記未知事例データの未知フィールドの値に依存する重みをフィールドごとに計算し、フィールドごとの重みを用いて前記類似度を計算する類似度計算手段とを備えることを特徴とする予測装置。

【請求項25】 前記類似度計算手段は、フィールドごとに前記既知事例データの集合を該フィールドの値により部分集合に分割し、前記未知事例データが属する部分集合の未知フィールドの値の分布が全体の未知フィールドの値の分布に近いほど小さくなり、該未知事例データが属する部分集合の未知フィールドの値の分布が該全体の未知フィールドの値の分布から遠いほど大きくなるような重みを計算することを特徴とする請求項24記載の予測装置。

【請求項26】 類似事例に基づく予測を行う予測装置であって、

類似事例抽出条件と既に得られている暫定的な類似事例データの集合を用いて、既知事例データを該暫定的な類似事例データの集合に加えるための類似度条件を計算する類似度条件計算手段と、

前記既知事例データと未知事例データの類似度を計算し、該類似度が前記類似度条件を満たすとき、該既知事例データを類似事例データとして出力し、該類似度が該類似度条件を満たさないことが確定したとき、計算を中止する条件付き類似度計算手段と、

前記条件付き類似度計算手段から出力された類似事例データを用いて、新たな類似事例データの集合を生成する生成手段とを備えることを特徴とする予測装置。

【請求項27】 前記生成手段は、前記条件付き類似度計算手段から出力された類似事例データを前記暫定的な類似事例データの集合に加え、前記類似事例抽出条件に合致するように余分な事例データを取り除いて、新しい類似事例データの集合を生成する類似事例集合更新手段と、該新しい類似事例データの集合を暫定的な類似事例

データの集合として記憶する類似事例集合記憶手段とを含み、与えられたすべての既知事例データの処理が終了したとき、該類似事例集合記憶手段に記憶された類似事例データの集合を出力することを特徴とする請求項26記載の予測装置。

【請求項28】 類似事例に基づく予測を行うコンピュータのためのプログラムを記録した記録媒体であって、1つ以上のフィールドからなる既知事例データの集合から未知事例データに類似する1つ以上の類似事例データを抽出するための類似事例抽出条件を自動的に決定するステップと、決定された類似事例抽出条件を用いて、前記1つ以上の類似事例データを抽出するステップと、前記1つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測し、予測値を生成するステップとを含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項29】 類似事例に基づく予測を行うコンピュータのためのプログラムを記録した記録媒体であって、1つ以上のフィールドからなる既知事例データの集合において、未知事例データの未知フィールドに対応するフィールドの値の分布を計算するステップと、前記分布に依存する重みをフィールドごとに計算するステップと、フィールドごとの重みを用いて類似度を計算するステップと、前記既知事例データの集合から、前記類似度に基づいて前記未知事例データに類似する1つ以上の類似事例データを抽出するステップと、前記1つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測し、予測値を生成するステップとを含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項30】 類似事例に基づく予測を行うコンピュータのためのプログラムを記録した記録媒体であって、類似事例抽出条件と既に得られている暫定的な類似事例データの集合を用いて、既知事例データを該暫定的な類似事例データの集合に加えるための類似度条件を計算するステップと、前記既知事例データと未知事例データの類似度を計算するステップと、前記類似度が前記類似度条件を満たすとき、該既知事例データを類似事例データとして決定するステップと、前記類似度が前記類似度条件を満たさないことが確定したとき、計算を中止するステップと、決定された類似事例データを用いて、新たな類似事例データの集合を生成するステップとを含む処理を前記コンピュータに実行させるためのプログラムを記録したコン

ピュータ読み取り可能な記録媒体。

【請求項31】 コンピュータを用いた類似事例に基づく予測方法であって、

1つ以上のフィールドからなる既知事例データの集合から未知事例データに類似する1つ以上の類似事例データを抽出するための類似事例抽出条件を自動的に決定し、決定された類似事例抽出条件を用いて、前記1つ以上の類似事例データを抽出し、

前記1つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測することを特徴とする予測方法。

【請求項32】 コンピュータを用いた類似事例に基づく予測方法であって、

1つ以上のフィールドからなる既知事例データの集合において、未知事例データの未知フィールドに対応するフィールドの値の分布を計算し、

前記分布に依存する重みをフィールドごとに計算し、

フィールドごとの重みを用いて類似度を計算し、

前記既知事例データの集合から、前記類似度に基づいて前記未知事例データに類似する1つ以上の類似事例データを抽出し、

前記1つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測することを特徴とする予測方法。

【請求項33】 コンピュータを用いた類似事例に基づく予測方法であって、

類似事例抽出条件と既に得られている暫定的な類似事例データの集合を用いて、既知事例データを該暫定的な類似事例データの集合に加えるための類似度条件を計算し、

前記既知事例データと未知事例データの類似度を計算し、

前記類似度が前記類似度条件を満たすとき、該既知事例データを類似事例データとして決定し、

前記類似度が前記類似度条件を満たさないことが確定したとき、計算を中止し、

決定された類似事例データを用いて、新たな類似事例データの集合を生成し、

与えられたすべての既知事例データの処理が終了したとき、生成された類似事例データの集合を用いて、前記未知事例データの未知フィールドの値を予測することを特徴とする予測方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、人工知能や統計処理等の分野において、与えられた未知事例に類似する類似事例を既知事例の中から抽出し、抽出された類似事例を用いて未知事例に関する予測を行う予測装置およびその方法に関する。

【0002】

【従来の技術】与えられた事例がどのクラス（カテゴリ）に属するかを決定するデータ分類方法は、人工知能や統計処理等の分野においてよく研究されている。事例（caseまたはinstance）とは、例えば、レコードのようなデータ構造に対応し、1つ以上のフィールドから構成される。そして、各フィールドには、属性データが記録される。

【0003】例えば、セールスキャンペーンにおいてダイレクトメールの送り先を決定する場合、送り先となる個人の情報を事例として扱い、多数の個人をフィールドの値により分類することが考えられる。このとき、各事例には、性別、年齢、住所、職業等の属性データのフィールドが設けられる。

【0004】また、既知事例の集合から未知事例の集合のクラスフィールドを予測する方法の1つとして、記憶に基づく推論（memory-based reasoning）、事例に基づく推論（case-based reasoning）等の類似事例に基づく予測が挙げられる。ここで、クラスフィールドは、予測対象となるフィールドを指し、既知事例は、クラスフィールドの値が既知の事例を表し、未知事例は、クラスフィールドの値が未知の事例を表す。

【0005】上述のダイレクトメールの例では、例えば、個人がダイレクトメールに対して応答を返したかどうかを表すフラグをクラスフィールドとして設定する。そして、既にダイレクトメールを送付した個人の集合を既知事例集合とし、これからダイレクトメールを送付しようとする個人の集合を未知事例集合として、未知事例集合のクラスフィールドの値（クラス値）を予測する。これにより、応答を返す可能性の高い個人を送り先として選択することが可能になる。

【0006】未知事例のクラスフィールドを予測するためには、まず、事例間の類似度を定義し、その未知事例に類似しているいくつかの事例を既知事例集合から取り出す。そして、それらの類似事例を基にして、未知事例のクラスフィールドを予測する。この予測方法は、ニューラルネットワークや決定木等を用いた学習やモデル化に基づく他の予測方法とは異なり、学習やモデル化のための時間が不要で、既知事例の増加に柔軟に対応できるという利点を持っている。

【0007】

【発明が解決しようとする課題】しかしながら、上述した従来の予測方法には、次のような問題がある。

【0008】ある未知事例に関する予測を行うためには、その未知事例と各既知事例との間の類似度を計算して、類似事例を抽出しなければならない。実は、類似事例に基づく予測に要する処理時間のほとんどが、この類似度計算の時間である。従来の予測方法では、類似度計算の時間を短縮するために、既知事例に対して一定の前処理を行う必要があった（毛利隆夫(Takao Mori), 「Nearest Neighbor法と記憶に基づく推論」, 人口知能学会

誌, Vol.12 No.2, p188-p195, March 1997. )。

【0009】この前処理では、既知事例集合をグループ分けして、未知事例との類似度が低いと思われるグループを類似度計算の対象から外す処理や、既知事例集合から不要な事例を削除する処理等が行われる。しかし、このような前処理を行うと、既知事例の増加に柔軟に対応することが難しくなる。

【0010】また、従来は、抽出される類似事例に関する条件として、類似度の高い上位20個の事例というようなデフォルトの条件をそのまま用いるか、あるいはユーザの手によって直接入力された条件を用いていた。しかし、どの条件が最適であるかは事例集合の性質および問題設定の目的によって大きく異なることが多く、適当な条件を指定しなければユーザが満足する結果が得られないことも多い。そこで、目的に合った条件を選ぶために、ユーザが条件を変更しながら類似事例の抽出を複数回実行し、その結果を評価して、最適な条件を選ぶことが多かった。

【0011】また、従来の予測方法では、類似事例を選択する際、クラス値に対する影響度に応じてフィールド毎に重みを設定することで、精度の向上が図られている。自動的な重み設定の方法としては、統計情報を用いたCross Category Feature (CCF) 法が存在する。しかし、CCF法によれば、既知事例集合のクラス値の分布に無関係に重みが設定され、クラス値の分布の変化による重みへの影響が十分ではない。このため、既知事例集合のクラス値の分布が大きく偏っている場合等には、十分な精度が出ないことが多かった。

【0012】このように、従来の予測方法では、既知事例集合に対する前処理なしに予測の高速化を行うことができず、既知事例の増加に対する柔軟性が損なわれるという問題がある。また、十分な予測精度を得るために類似事例の抽出を複数回行うので、ユーザの負担が大きくなり、実行回数に応じて計算時間が増大するという問題もある。さらに、既知事例集合のクラス値の分布が大きく偏っている場合等、問題の性質によっては、CCF法により十分な精度が得られないという問題もある。

【0013】本発明の課題は、類似事例に基づいて未知事例に関する予測を高速かつ高精度に行う予測装置およびその方法を提供することである。

【0014】

【課題を解決するための手段】図1は、本発明の予測装置の原理図である。

【0015】本発明の第1の原理による予測装置は、決定手段1と予測手段2を備え、類似事例に基づく予測を行う。

【0016】決定手段1は、1つ以上のフィールドからなる既知事例データの集合から未知事例データに類似する1つ以上の類似事例データを抽出するための類似事例抽出条件を自動的に決定する。

【0017】予測手段2は、決定手段1により決定された類似事例抽出条件を用いて、上記1つ以上の類似事例データを抽出し、それらの類似事例データを用いて、未知事例データの未知フィールドの値を予測し、予測値を出力する。

【0018】決定手段1は、既知事例データの集合から疑似的な未知事例データ（未知事例入力用事例データ）の集合を生成し、各未知事例入力用事例データのクラスフィールドが未知であるものとして予測を行う。そして、例えば、最良の結果が得られるような類似事例抽出条件を、最適な類似事例抽出条件として決定する。

【0019】また、予測手段2は、決定された類似事例抽出条件を用いて未知事例データの未知フィールド（クラスフィールド）の予測を行い、その類似事例抽出条件に応じた精度の予測値を得る。

【0020】このように、既知事例データの集合のみを用いてあらかじめ疑似的な予測処理を行うことで、最適な類似事例抽出条件が自動的に決定されるので、実際の未知事例データの予測は1回だけ行えばよい。したがって、従来のように未知事例データの予測を何度も繰り返さなくても、精度の高い予測値が得られ、処理が高速化される。

【0021】また、本発明の第2の原理による予測装置は、類似事例抽出手段3、予測結果生成手段4、および類似度計算手段5を備え、類似事例に基づく予測を行う。

【0022】類似事例抽出手段3は、1つ以上のフィールドからなる既知事例データの集合から、類似度に基づいて未知事例データに類似する1つ以上の類似事例データを抽出する。

【0023】予測結果生成手段4は、上記1つ以上の類似事例データを用いて、未知事例データの未知フィールドの値を予測し、予測値を出力する。

【0024】類似度計算手段5は、上記既知事例データの集合における上記未知フィールドの値の分布と上記未知事例データの未知フィールドの値に依存する重みをフィールドごとに計算し、フィールドごとの重みを用いて類似度を計算する。

【0025】類似度計算手段5が計算するフィールドごとの重みは、既知事例データの集合のクラス値の分布の影響を受けるため、それを用いて計算される類似度もそのクラス値の分布の影響を受ける。類似事例抽出手段3は、このような類似度に基づいて類似事例データを抽出し、予測結果生成手段4は、それらの類似事例データを用いて未知事例データのクラス値を予測する。

【0026】これにより、既知事例データの集合のクラス値の分布の偏りが大きい場合等の、従来のCCF法では十分な精度が得られないような問題においても、高精度な予測が行えるようになる。

【0027】また、本発明の第3の原理による予測装置

は、類似度条件計算手段6、条件付き類似度計算手段7、および生成手段8を備え、類似事例に基づく予測を行う。

【0028】類似度条件計算手段6は、類似事例抽出条件と既に得られている暫定的な類似事例データの集合を用いて、既知事例データを暫定的な類似事例データの集合に加えるための類似度条件を計算する。

【0029】条件付き類似度計算手段7は、既知事例データと未知事例データの類似度を計算し、類似度が類似度条件を満たすとき、その既知事例データを類似事例データとして出力し、類似度が類似度条件を満たさないことが確定したとき、計算を中止する。

【0030】生成手段8は、条件付き類似度計算手段7から出力された類似事例データを用いて、新たな類似事例データの集合を生成する。

【0031】条件付き類似度計算手段7は、類似度計算の途中で、類似度条件計算手段6が出力した類似度条件を満たされないことが確定した時点で、その既知事例データは類似事例データになり得ないと判断し、計算を中止する。そして、生成手段8は、類似度計算が中止されなかった既知事例データを用いて類似事例データの集合を生成し、予測装置は、その類似事例データの集合を用いて予測を行う。

【0032】このように、類似事例データの抽出において多大な時間を要する類似度計算を途中で打ち切ることによって、不要な計算時間が削減され、類似事例の抽出が高速化される。したがって、既知事例データの集合に対する前処理を行わなくても、高速な予測を行うことが可能になる。

【0033】例えば、図1の決定手段1と予測手段2は、それぞれ、後述する図2の類似事例抽出条件決定部12と予測部13に対応し、図1の類似事例抽出手段3と類似度計算手段5は、後述する図5の類似事例抽出部41に対応し、図1の予測結果生成手段4は、図5の予測結果生成部42に対応する。また、例えば、図1の類似度条件計算手段6と条件付き類似度計算手段7は、それぞれ、後述する図6の類似度条件計算部54と条件付き類似度計算部51に対応し、図1の生成手段8は、図6の類似事例集合更新部52と類似事例集合記憶部53に対応する。

【0034】

【発明の実施の形態】以下、図面を参照しながら、本発明の実施の形態を詳細に説明する。

【0035】本発明では、事例間の類似度計算の際に、ある既知事例が類似事例になり得ないことが分かった時点で計算を中止する。これにより、類似事例の抽出が高速化され、既知事例集合に対する前処理を行わなくても、予測が高速化される。

【0036】また、本発明では、従来、ユーザが明示的に処理を複数回実行することで得ていた最適な類似事例



抽出条件を、1回の実行で自動的に得ることができるようにする。ここでは、既知事例集合またはその部分集合をテスト用の疑似的な未知事例集合として用い、各未知事例のクラスフィールドが未知であるものとして予測を行う。そして、最良の結果が得られるような類似事例抽出条件を、最適な類似事例抽出条件として決定する。

【0037】このとき、条件を変更しながら類似事例の抽出を複数回行うことを避けるために、すべての類似事例抽出条件を包含する最も広い条件をあらかじめ計算しておき、その条件を用いて類似事例の抽出を1回だけ行う。その後、得られた類似事例を評価することで、最適な類似事例抽出条件を決定する。これにより、予測精度を損なうことなく、処理を高速化することができる。

【0038】また、本発明では、既知事例集合のクラス値の分布の影響を受け、クラス値の分布の変化による重みへの影響が従来のCCF法よりも大きくなるような影響度計算方法を用いる。これにより、既知事例集合のクラス値の分布の偏りが大きい場合等の、CCF法では十分な精度が得られないような問題においても、高精度な予測が行えるようになる。

【0039】図2は、本発明の予測装置の構成図である。図2の予測装置11は、例えば、コンピュータを用いて構成され、類似事例抽出条件決定部12と予測部13を備える。類似事例抽出条件決定部12および予測部13の機能は、例えば、コンピュータのメモリに格納されたプログラムを実行することで実現される。

【0040】類似事例抽出条件決定部12は、既知事例集合Aを用いて最適な類似事例抽出条件を決定し、それを出力する。予測部13は、類似事例抽出条件決定部12の出力を類似事例抽出条件として用い、既知事例集合Aから未知事例集合Uのクラスフィールドを予測して、予測結果を出力する。

【0041】図3は、図2の類似事例抽出条件決定部12の構成図である。図3の類似事例抽出条件決定部12は、入力用事例生成部21、類似事例抽出部22、類似事例削除部23、最適条件決定部24、条件出力部25、最大条件計算部26、および最大条件修正部27を備える。

【0042】入力用事例生成部21は、既知事例集合Aから既知事例入力用事例集合Bと未知事例入力用事例集合Cの2つの事例集合を生成する。条件出力部25は、あらかじめ保持している類似事例抽出条件の種類と最適条件選択範囲を出力する。最適条件選択範囲は、良好な予測を行うために最適な類似事例抽出条件を選択する際の条件の範囲を表す。

【0043】最大条件計算部26は、条件出力部25が出力するすべての条件を包含するような最も広い条件を求め、それを最大条件として出力する。最大条件修正部27は、既知事例入力用事例集合Bに未知事例入力用事例集合Cと重複した事例が含まれている場合に、最大条

件を修正して出力し、それ以外の場合は最大条件をそのまま出力する。

【0044】類似事例抽出部22は、入力用事例生成部21の出力B、Cを入力とし、最大条件修正部27が出力する条件に合致する類似事例集合を出力する。類似事例削除部23は、入力用事例生成部21の出力である既知事例入力用事例集合Bに重複した事例が含まれている場合に、類似事例抽出部22の出力である類似事例集合から事例を削除して、類似事例集合を修正する。

【0045】最適条件決定部24は、類似事例削除部23の出力を評価し、条件出力部25の出力である最適条件選択範囲の中から、良好な予測を行うために最適な類似事例抽出条件を決定し、それを出力する。

【0046】図4は、図3の最適条件決定部24の構成図である。図4の最適条件決定部24は、条件離散化部31、条件付き類似事例抽出部32、予測結果生成部33、条件評価部34、および最適条件選択部35を備える。

【0047】条件離散化部31は、条件出力部25の出力である最適条件選択範囲を離散化して出力する。条件付き類似事例抽出部32は、類似事例削除部23の出力から、条件離散化部31の出力の条件ごとに合致する類似事例集合を抽出する。

【0048】予測結果生成部33は、条件付き類似事例抽出部32の出力を用いて、条件離散化部31の出力の条件ごとに予測を行う。条件評価部34は、予測結果生成部33の出力である予測結果を評価し、条件離散化部31の出力の条件ごとに評価値を求める。最適条件選択部35は、条件評価部34の出力である評価値に基づいて、良好な予測を行うために最適な類似事例抽出条件を選択する。

【0049】このような構成によれば、類似事例抽出部22による類似事例の抽出は1回行われるだけであり、最適な類似事例抽出条件は最適条件決定部24により自動的に決定される。類似事例に基づく予測では、計算時間のほとんどが類似事例抽出の際の類似度計算に費やされるため、類似事例抽出を1回で済ますことにより、ユーザが明示的に複数回実行する場合と比べて、計算時間を大幅に短縮できる。

【0050】ところで、図3の条件出力部25は、類似事例抽出条件の種類として、例えば、1つの未知事例に対する類似事例の数がkであるという条件を出力し、最適条件選択範囲としてkの値の検査範囲を出力する。kの値の範囲としては、例えば、 $1 \leq k \leq k_1$ なる整数が指定される。

【0051】この場合、最大条件計算部26は、類似事例数がk1であるという条件を最大条件として出力する。k1の値は、ユーザが指定することもでき、 $k_1 = (\text{既知事例集合Aの事例数の平方根})$ のように、システムが自動的に設定することもできる。



【0052】また、条件出力部25は、類似事例抽出条件の種類として、類似度が閾値 $s$ 以上という条件を出力し、最適条件選択範囲として閾値 $s$ の値の検査範囲を出力することもできる。 $s$ の値の範囲としては、例えば、 $s_1 \leq s \leq s_2$ が指定される。この場合、最大条件計算部26は、類似度が $s_1$ 以上という条件を最大条件として出力する。

【0053】 $s_1$ 、 $s_2$ の値は、それぞれ、ユーザが指定することもでき、システムが自動的に設定することもできる。後者の場合、例えば、 $\alpha = 1$ 、 $\beta = 100$ のよ

$$a = 0 \quad \text{or} \quad 1/k_1 \leq a \leq 1$$

$$b = 0 \quad \text{or} \quad s_1 \leq b \leq s_2$$

ただし、 $a > 0$ もしくは $b > 0$ であるものとする。 $k_1$ 、 $s_1$ 、および $s_2$ の値は、上述したような方法で設  
 $s \geq s_1$ もしくは $k = k_1$

次に、図5は、図2の予測部13の構成図である。図5の予測部13は、類似事例抽出部41と予測結果生成部42を備える。類似事例抽出部41は、既知事例集合Aと未知事例集合Uを入力とし、類似事例抽出条件決定部12が出力する条件に合致する類似事例集合を出力する。予測結果生成部42は、類似事例抽出部41の出力を用いて予測を行い、予測結果を出力する。

【0055】図6は、図5の類似事例抽出部41の構成図である。図6の類似事例抽出部41は、条件付き類似度計算部51、類似事例集合更新部52、類似事例集合記憶部53、および類似度条件計算部54を備え、未知事例集合Uの各事例ごとに、既知事例集合Aから類似事例集合を抽出する。

【0056】条件付き類似度計算部51は、既知事例集合Aから既知事例を1つずつ取り出して、既知事例と未知事例の類似度を計算し、与えられた類似度条件を満たす既知事例とその類似度を出力する。ただし、与えられた類似度条件を満たさないことが分かった時点で、計算を中止する。

【0057】類似事例集合更新部52は、既に得られている暫定的な類似事例集合に条件付き類似度計算部51の出力を加え、類似事例抽出条件に合致するように、余分な事例を取り除き、新しい類似事例集合を出力する。類似事例集合記憶部53は、類似事例集合更新部52の出力を現在の暫定的な類似事例集合として記憶し、それを類似事例集合更新部52および類似度条件計算部54に出力する。

【0058】類似度条件計算部54は、類似事例集合記憶部53の内容と類似事例抽出条件から、ある事例が新しく類似事例集合に加わるための必要十分条件である類似度条件を計算し、それを条件付き類似度計算部51に出力する。

【0059】このように、条件付き類似度計算部51が類似度条件に基づいて不要な類似度計算を途中で中止することにより、既知事例集合からの類似事例抽出が高速

うにあらかじめ設定されたパラメータを用いて、 $s_1 = \alpha / (\text{既知事例集合Aのフィールド数})$ 、 $s_2 = \beta / (\text{既知事例集合Aのフィールド数})$ のように設定することができる。

【0054】また、条件出力部25は、類似事例抽出条件の種類として、類似事例数 $k$ と類似度 $s$ を含む条件を出力することもできる。例えば、 $a k + b / s \leq 1$ という類似事例抽出条件を出力する場合、次のようなパラメータ $a$ 、 $b$ の値の検査範囲を最適条件選択範囲として出力する。

$$(1)$$

$$(2)$$

定することができる。この場合、最大条件計算部26は、次のような条件を最大条件として出力する。

$$(3)$$

化され、類似事例に基づく予測を高速化することができる。この類似事例抽出部41の構成を図3の類似事例抽出部22に実装して、類似事例抽出条件決定部12の処理をより高速化することもできる。

【0060】次に、図7から図49までを参照しながら、上述した予測装置11の動作についてより詳細に説明する。

【0061】図7は、図3の入力用事例生成部21の第1の例を示している。図7の入力用事例生成部21は分割部61を備え、既知事例集合Aを2つに分割して、一方を既知事例入力用事例集合Bとして出力し、もう一方を未知事例入力用事例集合Cとして出力する。分割部61における分割方法としては、例えば、ランダムサンプリング等が考えられる。

【0062】このとき、最適な類似事例抽出条件が既知事例集合Aと既知事例入力用事例集合Bの間で大きく異なるようにするために、未知事例入力用事例集合Cの事例数が既知事例集合Aの事例数に比べて十分小さいことが望ましい。この構成では、既知事例入力用事例集合Bには、未知事例と重複する事例は含まれないため、最大条件修正部27および類似事例削除部23は何も行わず、入力をそのまま出力する。

【0063】図8は、図3の入力用事例生成部21の第2の例を示している。図8の入力用事例生成部21は事例削除部62を備え、既知事例集合Aをそのまま未知事例入力用事例集合Cとして出力する。事例削除部62は、未知事例入力用事例集合Cの各未知事例ごとに、既知事例集合Aからその未知事例を削除する。そして、未知事例ごとに異なる事例集合を生成し、それを既知事例入力用事例集合Bとする。

【0064】この場合、既知事例入力用事例集合Bは、未知事例を含まず、かつ、既知事例集合Aに最も近い事例集合であるといえる。この構成においても、既知事例入力用事例集合Bには、未知事例と重複する事例は含まれないため、最大条件修正部27および類似事例削除部

23は何も行わず、入力をそのまま出力する。

【0065】図9は、図3の入力用事例生成部21の第3の例を示している。図9の入力用事例生成部21は、既知事例集合Aをそのまま未知事例入力用事例集合Cとして出力し、同一の既知事例集合Aをそのまま既知事例入力用事例集合Bとしても出力する。

【0066】この構成では、既知事例入力用事例集合Bに、未知事例と重複する事例が含まれるため、最大条件修正部27は、類似事例を1つ余計に抽出できるように最大条件を修正し、類似事例削除部23は、各未知事例の類似事例集合からその未知事例と重複する事例を削除する。

【0067】図10は、図3の入力用事例生成部21の第4の例を示している。図10の入力用事例生成部21は、図8の入力用事例生成部21にサンプリング部63を付加した構成を持つ。サンプリング部63は、既知事例集合Aの事例数が一定数（例えば、1000）よりも多い場合に、ランダムサンプリング等により事例をサンプリングする。そして、既知事例集合Aの一部を未知事例入力用事例集合Cとして出力する。

【0068】事例削除部62は、サンプリング部63が出力した未知事例入力用事例集合Cの各未知事例ごとに、既知事例集合Aからその未知事例を削除する。そして、未知事例ごとに異なる事例集合を生成し、それを既知事例入力用事例集合Bとする。

【0069】この構成によれば、既知事例集合Aが比較的大きな場合に、未知事例入力用事例集合Cの大きさを限定することができ、後続する類似事例抽出部22の処理を高速化することができる。

【0070】図11は、図3の入力用事例生成部21の第5の例を示している。図11の入力用事例生成部21は、図9の入力用事例生成部21にサンプリング部63を付加した構成を持つ。サンプリング部63の動作は、図10の場合と同様である。

【0071】この構成では、既知事例入力用事例集合Bに、未知事例と重複する事例が含まれるため、最大条件修正部27は、図9の場合と同様に最大条件を修正し、類似事例削除部23は、各未知事例の類似事例集合からその未知事例と重複する事例を削除する。

【0072】ここで、具体的な事例集合を用いて入力用事例生成部21の処理を説明する。例えば、ダイレクトメールの送り先からの応答の有無を予測するために、図12のような既知事例集合Aを与えられたものとする。

【0073】図12において、1つの行が1つの事例のレコードに対応し、“氏名”は、事例を識別するための個人名を表す。この既知事例集合には、“A”、“B”、“C”、“D”、“E”、“F”、および“G”の7人の事例が含まれている。各事例は、“年齢”、“性別”、“職業”、“結婚”、および“応答”のフィールドを含んでおり、このうち、“応答”がクラ

スフィールドに対応する。

【0074】例えば、“A”さんの事例には、年齢が“30歳”で、性別が“男”で、職業が“公務員”で、結婚については“既婚”で、ダイレクトメールへの応答が“あり”であることが記録されている。また、“C”さんの事例には、年齢が“40歳”で、性別が“女”で、職業が“無職”で、結婚については“既婚”で、ダイレクトメールへの応答が“なし”であることが記録されている。

【0075】ここで、図11に示した入力用事例生成部21を用いてサンプリングを行い、図13のような未知事例入力用事例集合Cが得られたとする。図13の未知事例入力用事例集合には、“A”、“C”、“E”、および“G”の4人の事例が含まれている。この場合、既知事例入力用事例集合Bとしては、図12の事例集合がそのまま用いられる。

【0076】図3の類似事例抽出部22は、記憶に基づく推論もしくは事例に基づく推論により、入力用事例生成部21が出力した集合から類似事例を抽出する。記憶に基づく推論および事例に基づく推論において、類似事例抽出部22は基本的に同じ処理を行う。以下に類似事例抽出部22による類似度計算の一例を示す。

【0077】まず、事例の各フィールドごとに、クラス値の決定に対する影響度を計算する。ここでは、統計情報を基にした影響度計算方法として広く知られているCCF法を用いることにする。この方法では、クラスフィールドも含めて、各フィールドごとに、フィールド値がいくつかの領域に分けられる。

【0078】例えば、図12の“性別”、“職業”、“結婚”、および“応答”のように、カテゴリを表すカテゴリ値フィールドであれば、そのフィールドが表現し得るカテゴリごとにフィールド値を分類し、図12の“年齢”のように数値を表す数値フィールドであれば、数値の区間ごとにフィールド値を分類する。ただし、“年齢”のフィールドは、離散的な数値を表すため、カテゴリ値フィールドとして扱うことも可能である。

【0079】今、既知事例集合の事例番号を*i*とし、フィールド番号を*j*とすると、フィールド*j*の値が領域*v*に含まれているときに、クラスフィールドの値が領域*c*に含まれている条件付き確率を、 $p(j, v, c)$ のように表示することができる。このとき、フィールド*j*の重み  $w(j, v)$  は、次式で与えられる。

【0080】

【数1】

$$w(j, v) = \sum_c p(j, v, c)^2 \quad (4)$$

【0081】次に、2つのフィールド値間の距離を定義する。ここでは、一例として、最も単純なフィールド間距離の1つである次式のような距離  $d(j)$  を用いる。

【0082】

【数2】

$$d(j) = \begin{cases} \frac{\text{field値間の差}}{\text{標準偏差}} & (\text{数値 field}) \\ 0 & (\text{field値が一致}) \\ 1 & (\text{field値が不一致}) \end{cases} \quad (\text{カテゴリ値 field}) \quad (5)$$

【0083】(5)式によれば、数値フィールドの場合は、フィールド値の分布から求められた標準偏差を分母とし、2つのフィールド値の差を分子とする値が、距離 $d(j)$ として用いられる。また、カテゴリ値フィールドの場合は、2つのフィールド値が一致したとき $d(j) = 0$ と定義され、それが一致しなかったとき $d(j) = 1$ と定義される。

【0084】次に、2つの事例間の類似度を定義する。ここでは、一例として、最も単純な類似度の1つである次式のような類似度 $S$ を用いる。

【0085】

【数3】

$$S = 1 / \sqrt{\sum_j w(j, v(j)) \times d(j)^2} \quad (6)$$

【0086】ただし、 $v(j)$ は、未知事例のフィールド $j$ の値が属している領域を表す。こうして、(4)式、(5)式、および(6)式により事例間類似度が定義されたので、各未知事例についてすべての既知事例との類似度を計算し、最大条件修正部27の出力条件に合致するような既知事例を選択することで、未知事例ごとに類似事例を抽出することができる。また、このような類似度計算は、図6の条件付き類似度計算部51にも適用することができる。

【0087】例えば、条件出力部25が類似事例の数 $k$ を類似事例抽出条件として出力し、 $k \leq 5$ を最適条件選択範囲として出力したとすると、最大条件計算部26は、 $k = 5$ を最大条件として出力する。このとき、最大条件修正部27は、類似事例を1つ余計に抽出できるように、最大条件を $k = 6$ に修正し、類似事例抽出部22は、修正された最大条件に従って類似事例を抽出する。

【0088】このとき、図12および図13に示した事例集合からは、例えば、図14、15、16、および1

$$s(i) = s1 + i * (s2 - s1) / 100 \quad (7)$$

ただし、 $i$ は、 $0 \leq i \leq 100$ なる整数である。(7)式により、 $s$ は、 $s(0)$ から $s(100)$ までの101点に離散化され、これに対応して101個の類似事例抽出条件が生成される。

【0093】また、類似事例数もしくは類似事例数と類似度を含む条件が類似事例抽出条件として与えられた場合も、同様の方法により条件を離散化することができる。例えば、類似事例数が $k$ という類似事例抽出条件と、 $k \leq 5$ という最適条件選択範囲が与えられた場合、類似事例抽出条件は、“類似事例数=1”、“類似事例

7に示すような類似事例集合が生成される。図14、15、16、および17の類似事例集合は、それぞれ、テスト用の未知事例“A”、“C”、“E”、および“G”に類似する事例の集合を表す。

【0089】ここでは、いずれの類似事例集合も、修正された最大条件に対応する6つの類似事例から構成され、それらの類似事例は類似度の大きい順に並べられている。各類似事例集合は、比較対象の未知事例と同じ類似事例を含んでおり、その類似事例の類似度は“\*\*\*”で表されている。

【0090】類似事例削除部23は、これらの類似事例集合から対応する未知事例と重複する事例を削除して、類似事例集合を修正する。図14、15、16、および17の類似事例集合から重複事例がそれぞれ削除された結果、図18、19、20、および21に示すような類似事例集合が生成される。例えば、図14の類似事例集合からは事例“A”が削除されて、図18の類似事例集合が生成されている。修正された類似事例集合は、それぞれ、修正される前の最大条件に対応する5つの類似事例から構成されている。

【0091】こうして修正された類似事例集合は、図4の最適条件決定部24に入力され、最適な類似事例抽出条件を決定するために利用される。図4の条件離散化部31は、連続な条件をあらかじめ決められた方法で離散化する。最も簡単な離散化方法としては、あらかじめ離散化数を決めておき、等間隔で条件を離散化する方法が挙げられる。

【0092】例えば、類似度が $s$ 以上という類似事例抽出条件と、 $s1 \leq s \leq s2$ という最適条件選択範囲が与えられた場合、次式のような類似度 $s(i)$ を用いて、条件を離散化することができる。

数=2”、“類似事例数=3”、“類似事例数=4”、および“類似事例数=5”の5つの条件に離散化される。

【0094】次に、条件付き類似事例抽出部32は、入力された類似事例集合から離散化された各条件に合致する類似事例を抽出し、条件ごとに異なる類似事例集合を出力する。

【0095】例えば、図18、19、20、および21に示した類似事例集合が入力された場合、“類似事例数=1”という条件に従うと、各類似事例集合から類似度

が最大の類似事例が1つずつ抽出され、図22、23、24、および25に示すような類似事例集合が生成される。

【0096】また、“類似事例数=2”という条件に従うと、各類似事例集合から、類似度の大きなものから順に類似事例が2つずつ抽出され、図26、27、28、および29に示すような類似事例集合が生成される。

【0097】また、“類似事例数=3”という条件に従うと、各類似事例集合から、類似度の大きなものから順に類似事例が3つずつ抽出され、図30、31、32、および33に示すような類似事例集合が生成される。

【0098】また、“類似事例数=4”という条件に従うと、各類似事例集合から、類似度の大きなものから順に類似事例が4つずつ抽出され、図34、35、36、および37に示すような類似事例集合が生成される。

【0099】また、“類似事例数=5”という条件は類似事例抽出部22が用いた条件と一致するため、この条件に従うと、図18、19、20、および21の類似事例集合がそのまま出力される。

【0100】次に、予測結果生成部33は、条件付き類似事例抽出部32が出力した類似事例集合を用いて、条件ごとにクラスフィールドの予測を行う。予測結果生成部33が記憶に基づく推論により予測を行う場合、類似事例集合および各類似事例に付随する類似度を用いて、以下のような処理を行う。

【0101】まず、クラス値がカテゴリ値の場合、クラス値ごとに、そのクラス値 $c$ を有する類似事例の事例間類似度を足し合わせることで、クラス値ごとの類似度 $T(c)$ を計算する。そして、類似度 $T(c)$ が最大となるクラス値を予測クラス値(予測値) $c(\text{predict})$ として、予測値のもらしさを表す確信度 $P$ を、次式により定義する。

【0102】

【数4】

$$P = \frac{T(c(\text{predict}))}{\sum_c T(c)} \quad (8)$$

【0103】(8)式の分子の $T(c(\text{predict}))$ は、予測値 $c(\text{predict})$ に対応する類似度を表し、(8)式の分母は、すべてのクラス値についての $T(c)$ の総和を表す。したがって、確信度 $P$ は、最大値が1であるような正の数値である。このような予測値の計算方法のほかに、類似事例集合の事例中に最も多く現れるクラス値を予測値としてもよい。

【0104】また、クラス値が連続値の場合は、予測値および確信度を、例えば、次式により定義することができる。

【0105】

【数5】

$$c(\text{predict}) = \frac{\sum_i S(i)c(i)}{\sum_i S(i)} \quad (9)$$

$$P = \frac{1}{\frac{\sum_i S(i)(c(i) - c(\text{predict}))^2}{\sigma(c)^2 \sum_i S(i)} + 1} \quad (10)$$

【0106】ここで、 $n$ は類似事例集合に含まれる事例数を表し、 $S(i)$ は事例 $i$  ( $i=1, 2, \dots, n$ )の事例間類似度を表し、 $c(i)$ は事例 $i$ のクラス値を表し、 $\sigma(c)$ はクラス値 $c(i)$ の分布における標準偏差を表す。こうして算出された予測値および確信度は、予測結果として出力される。

【0107】例えば、図22から図37に示した条件ごとの類似事例集合の場合、クラスフィールド“応答”はカテゴリフィールドであるから、クラス値ごとの類似度 $T(c)$ が計算され、最大の $T(c)$ に対応するクラス値が予測値として求められる。その結果、図38に示すような予測値が得られる。

【0108】図38のマトリクスにおいて、各行は、未知事例入力用事例集合 $C$ の1つの未知事例に対応し、各列は、1つの類似事例抽出条件(類似事例数)に対応する。したがって、マトリクスの1つの要素(セル)は、列に対応する類似事例抽出条件に従って抽出された類似事例集合を用いて、行に対応する未知事例のクラスフィールド“応答”の値を予測したときの予測値を表す。

【0109】例えば、“類似事例数=1”の場合は、図22、23、24、および25に示したように、各類似事例集合は1つの事例しか含まないため、その事例の“応答”の値がそのまま予測値として採用される。

【0110】また、“類似事例数=2”の場合は、図26、27、28、および29に示したように、各類似事例集合は2つの事例を含む。例えば、未知事例“A”に対応する図26の類似事例集合では、クラス値“なし”を有する事例とクラス値“あり”を有する事例が1つずつ含まれており、前者の方が類似度が大きいため、“なし”が予測値として採用される。

【0111】また、“類似事例数=5”の場合は、図18、19、20、および21に示したように、各類似事例集合は5つの事例を含む。例えば、未知事例“A”に対応する図18の類似事例集合では、クラス値“なし”を有する3つの事例とクラス値“あり”を有する2つの事例が含まれており、前者の3つの事例の類似度の合計は9.5であり、後者の2つの事例の類似度の合計は7である。したがって、“なし”の類似度の方が“あり”の類似度より大きいため、“なし”が予測値として採用される。

【0112】次に、予測結果生成部33が事例に基づく推論により予測を行う場合、予測結果生成部33は図39に示すような構成を持つ。事例に基づく推論の予測結果生成部と記憶に基づく推論の予測結果生成部との相違

点は、前者が予測修正部72を備えていることである。

【0113】図39において、一次予測生成部71は、上述した記憶に基づく推論の予測結果生成部に対応し、予測修正部72は、一次予測生成部71の出力である一次予測結果を先見的知識等により修正して、予測結果生成部33の出力を生成する。先見的知識は、過去の予測結果に基づく経験則を表す。

【0114】上述した記憶に基づく推論または事例に基づく推論を用いた予測結果生成部33は、図5に示した予測部13内の予測結果生成部42としても用いることができる。

【0115】次に、条件評価部34は、未知事例入力用事例集合Cを用いて条件ごとの予測結果を評価し、各条件に対する評価値を生成する。テスト用の未知事例のクラス値は以下では、評価値の値が大きいほど良好な評価結果を表すものとして、条件評価部34のいくつかの実施形態について説明する。

【0116】まず、クラス値がカテゴリ値である場合、条件評価部34は、予測結果生成部33から出力された予測値と未知事例のクラス値（真のクラス値）を比較し、予測値と一致した事例の数を評価値とする。

【0117】この評価方法により、図38に示した予測結果を評価すると、図40に示すような評価値が得られる。ここでは、図13の未知事例入力用事例集合の各事例のクラスフィールド“応答”の値と図38の対応する予測値が比較される。類似事例数＝1、2、4、5の場合は、未知事例“C”および“E”についてのみ両者が一致しているので、評価値は“2”となり、類似事例数＝3の場合は、未知事例“A”、“C”、および“E”について両者が一致しているので、評価値は“3”となる。

【0118】また、予測値と一致した事例の数から一致しなかった事例の数を差し引いた結果を評価値としてもよい。例えば、未知事例入力用事例集合Cの未知事例数がmであり、一致した事例数がm1である場合、一致しなかった事例数はm-m1となり、評価値はm1-(m-m1)=2×m1-mとなる。

【0119】また、クラス値が連続値である場合、条件評価部34は、予測値と未知事例のクラス値との差の絶対値の平均をとり、得られた平均値に-1を掛けて評価値とする。この場合、評価値は負の値となる。

$$(\text{類推実行時間}) = (\text{既知事例数}) \times (\text{フィールド数}) + \alpha \times (\text{類似事例数})$$

$$(11)$$

$$(\text{評価値}) = (\text{実行時間を加味する前の評価値}) - \beta \times (\text{類推実行時間})$$

$$(12)$$

(11)式において、類似事例数が類似事例抽出条件のみから定まらない場合は、条件付き類似事例抽出部32から出力された類似事例集合の事例数の平均をとればよい。また、パラメータ $\alpha$ は、類似事例抽出条件を基に決められ、パラメータ $\beta$ は、ユーザが実行時間をどれほど

【0120】これらの評価方法では、予測結果生成部33の出力のうち予測値のみを用いて評価値が生成されているが、さらに確信度を加味して評価値を生成することもできる。まず、クラス値がカテゴリ値の場合、条件評価部34は、評価値の初期値を0とおき、各未知事例に対して、予測値と未知事例のクラス値が一致した場合にはその予測値の確信度を評価値に加え、両者が一致しなかった場合にはその予測値の確信度を評価値から減ずる。これにより、最終的に確信度を加味した評価値が得られる。

【0121】また、予測値と未知事例のクラス値を用いてクラスフィールドの重みを求め、その重みを加味して評価値を生成することもできる。例えば、クラスフィールドが“+”と“-”の2値のいずれかをとり、ユーザは、真のクラス値が“+”であるときに“-”という予測値を得ることはできるだけ避けたいと考えているとする。

【0122】このとき、ユーザは、避けたい予測パターンとして（予測値，真のクラス値）＝（-，+）を指定し、条件評価部34は、予測値と未知事例の真のクラス値の組合せに応じて、以下のような重みwを設定する。

$$(+, +) \text{ の場合 } w = 1.0$$

$$(+, -) \text{ の場合 } w = 1.0$$

$$(-, +) \text{ の場合 } w = 2.0$$

$$(-, -) \text{ の場合 } w = 1.0$$

そして、評価値の初期値を0とおき、各未知事例に対して、予測値と未知事例のクラス値が一致した場合には両者の組合せから得られる重みを評価値に加え、両者が一致しなかった場合にはその重みを評価値から減ずる。これにより、最終的にクラスフィールドの重みを加味した評価値が得られる。

【0123】また、類似事例抽出条件によっては、抽出される類似事例の数が多くなると実行時間が長くなる場合がある。このような場合、類似事例抽出条件から実行時間を類推し、実行時間の長さに応じた値を評価値から減ずることもできる。これにより、上述した評価値がほとんど変わらないいくつかの条件の中から、長い実行時間を要する条件を選択してしまうことが回避される。類推実行時間と実行時間を加味した評価値は、例えば、次式により与えられる。

重要と認識するかを基に決められる。

【0124】次に、最適条件選択部35は、条件評価部34から出力された評価値に基づいて、離散化された類似事例抽出条件の中から最適な条件を選択する。例えば、類似事例抽出条件として類似事例数が指定され、横

軸を類似事例数とし縦軸を評価値として図41に示すような評価値の分布が得られたものとする。

【0125】このとき、最適条件選択部35は、評価値が最良となる条件を最適条件として選択するため、円で囲まれた最大値に対応する類似事例数が選択される。例えば、図40に示した評価値の場合、最大値“3”に対応する“類似事例数=3”が最適条件として選択される。

【0126】また、最適条件選択部35は、評価値の移動平均を用いて最適な条件を選択することもできる。この場合、最適条件選択部35は、条件ごとに適当な領域を定めて、その領域内に含まれる複数の条件の評価値について平均をとる。そして、得られた平均値に対応する条件の新たな評価値とし、その評価値が最大となる条件を最適条件として選択する。このような選択方法によれば、ノイズにより評価値が細かく振動しているような場合でも、それが滑らかに平均化されるため、より安定な条件を選択することができる。

【0127】例えば、図41の評価値分布の場合、移動平均により図42のような平均評価値が得られる。ここでは、基準となる条件の類似事例数kに対応する移動平均の領域を、MAX(1, k-2)以上MIN(k\_max, k+2)以下としている。ただし、MAX()、MIN()は、それぞれ()内の数値の大きい方、小さい方をとるものとし、k\_maxは類似事例数の最大値とする。

【0128】そして、kの値を1からk\_maxまでインクリメントしながら、この領域内の類似事例数に対する評価値の平均値を求めていくと、破線で示した平均評価値の分布が得られる。このとき、円で囲まれた平均評価値に対応する類似事例数が最適条件として選択される。

【0129】また、最適条件選択部35は、評価値の近似関数を用いて最適な条件を選択することもできる。この場合、最適条件選択部35は、すべての条件に渡る評価値の分布を適当な関数で近似し、その関数の値が最大となる条件を選択する。この近似関数としては、例えば、条件を記述するパラメータのn次多項式が用いられる。nは、1以上の整数であるが、実験的には、4、5等の値が適当であることが分かっている。

【0130】例えば、図41の評価値分布の場合、近似関数により図43のような評価値が得られる。ここでは、類似事例数kの2次多項式を近似関数f(k)として用い、次のような値が最小となるように各項の係数を定めた。

【0131】

【数6】

$$\sum_i (y_i - f(k_i))^2 \quad (13)$$

【0132】ただし、k\_i と y\_i は、それぞれ、図41

の類似事例数と対応する評価値を表す。その結果、図43に破線で示した近似関数が得られ、円で囲まれた関数値に対応する類似事例数が最適条件として選択される。

【0133】上述した各選択方法は、類似事例数だけでなく、類似度等の他の任意の類似事例抽出条件にも同様に適用される。類似事例数と類似度のよう、異なる2つ以上の変数を含む類似事例抽出条件が与えられた場合は、変数ごとに独立に近似関数を求める等の方法により、図43の選択方法を採用することが可能である。

【0134】こうして決定された最適な類似事例抽出条件は、図2に示したように、予測部13に入力され、予測部13は、入力された条件を用いて未知事例集合Uに関する予測を行う。このように、図2の構成によれば、既知事例集合Aと未知事例集合Dを指定することで、自動的に最適な類似事例抽出条件が決定され、その条件を用いた予測が行われる。

【0135】ところで、予測部13に含まれる図6の条件付き類似度計算部51は、類似度計算において、既存のCCF法以外に他の任意の影響度計算方法を用いることができる。図3の類似事例抽出部22についても同様である。以下では、既知事例集合のクラス値の分布の影響を受け、クラス値の分布の変化による重みへの影響がCCF法よりも大きくなるような重み付けを用いる計算方法について説明する。

【0136】まず、フィールドjの値が領域vに含まれているときに、クラス値が領域cに含まれている条件付き確率をp(j, v, c)とし、クラス値が領域cに含まれる確率をp(c)とし、クラス値の数をN(c)とする。例えば、2値のクラスフィールドについては、N(c)=2である。このとき、フィールドjの重みw(j, v)は、(4)式の代わりに次式で与えられる。

【0137】

【数7】

$$q(j, v, c) = p(j, v, c) / p(c) \quad (14)$$

$$w(j, v) = \frac{\sum_c \left| \frac{q(j, v, c)}{\sum_d q(j, v, d)} \frac{1}{N(c)} \right|}{2 - \frac{2}{N(c)}} \quad (15)$$

【0138】ここで、p(j, v, c)は、既知事例集合をフィールドjの値により部分集合に分割した場合の領域vに対応する部分集合内のクラス値の分布に対応し、p(c)は全体のクラス値の分布に対応する。したがって、(14)式のq(j, v, c)は、既知事例集合の部分集合のクラス値の分布と全体のクラス値の分布の比を表している。

【0139】p(j, v, c)がp(c)と等しいとき、q(j, v, c)=1となり、(15)式の分子は0となる。したがって、このとき、重みw(j, v)は最小値0をとる。また、特定のクラス値cについてのみ



$p(j, v, c) = 1$  となり、 $c$  が他の値のとき  $p(j, v, c) = 0$  となるような場合は、(15) 式の分子は最大値  $2 - 2/N(c)$  をとる。(15) 式の分母はこの最大値に一致しているため、このとき、重み  $w(j, v)$  は最大値 1 をとる。

【0140】言い換えれば、部分集合のクラス値の分布が全体のクラス値の分布に近いほど、対応するフィールドの影響度は小さく、部分集合のクラス値の分布が全体のクラス値の分布から遠いほど、対応するフィールドの影響度は大きくなる。この影響度計算方法による重み  $w(j, v)$  を (6) 式の  $w(j, v(j))$  として用いれば、既知事例集合のクラス値の分布を類似度計算に反映させることができ、既知事例集合のクラス値の分布が偏っているような場合でも、高い精度の予測が可能になる。

【0141】次に、図6の類似事例抽出部41の動作について詳細に説明する。ここでは、類似事例数が  $k$  であるという類似事例抽出条件が与えられたものとする。条件付き類似度計算部51は、類似度計算の過程で、上述した(6)式の平方根の中の総和を求めるとき、フィールド1に対応する項から順番に加算していく。

【0142】この加算において、総和は単調に増加し、それに伴って類似度  $S$  は単調に減少する。したがって、既知事例を新しく類似事例集合に加えるための類似度条件が " $S \geq S1$ " であるとする、 $S < S1$  となった時点で、それ以上加算を続けてもその既知事例は類似事例になり得ないことが分かる。

【0143】そこで、条件付き類似度計算部51は、一定間隔で  $S \geq S1$  であるかどうかをチェックし、この条件が満たされないときには類似度計算を中止して、次の既知事例の類似度計算を開始する。そして、類似度条件を満たす既知事例を新たな類似事例として出力する。

【0144】類似事例集合更新部52は、類似事例集合記憶部53から現在の類似事例集合を取り出し、類似事例抽出条件に従って、条件付き類似度計算部51から出力された類似事例を類似事例集合に加える。このとき、新しい類似事例集合の事例数が  $k$  以下であれば、その類似事例集合を更新結果として出力し、新しい類似事例集合の事例数が  $k+1$  であれば、類似度が最も小さい事例を削除し、得られた類似事例集合を更新結果として出力する。

【0145】類似事例集合記憶部53は、類似事例集合更新部52から出力された類似事例集合を記憶する。ただし、初期状態においては、類似事例集合は空集合である。類似度条件計算部54は、類似事例集合記憶部53内の類似事例集合の事例数が  $k$  であるとき、" $S \geq S1$ " という類似度条件を出力する。 $S1$  としては、例えば、類似事例集合内の事例の類似度の最小値が用いられる。

【0146】また、類似事例集合の事例数が  $k$  未満であ

るときは、"条件なし" を出力する。この場合、条件付き類似度計算部51は、類似度計算を途中で打ち切ることなく、どんな類似度の事例であっても類似事例として出力する。

【0147】このように、類似事例抽出条件と既に得られている類似事例の類似度に応じて類似度計算を中止するための条件を決定し、その条件に従って計算を中止することで、類似事例抽出の効率が向上する。

【0148】次に、図2の予測装置11の主要部により行われる処理の例について、フローチャートを参照しながらより詳細に説明する。

【0149】図44は、図8の事例削除部62の処理のフローチャートである。事例削除部62は、まず、未知事例入力用事例集合Cの事例の番号を表す制御変数  $I$  を1とおき(ステップS1)、 $I$  を未知事例入力用事例集合Cの事例数  $|C|$  と比較する(ステップS2)。ここでは、未知事例入力用事例集合Cは既知事例集合Aに一致している。

【0150】 $I \leq |C|$  であれば、集合Cの  $I$  番目の事例  $C[I]$  を既知事例集合Aから削除して、事例  $C[I]$  に対応する既知事例入力用事例集合  $B[I]$  を生成し、 $I = I + 1$  において(ステップS3)、ステップS2以降の処理を繰り返す。そして、ステップS2において  $I > |C|$  となると、得られた集合  $B[I]$  を既知事例入力用事例集合Bとして出力し、処理を終了する。

【0151】次に、図45は、図3の類似事例削除部23の処理のフローチャートである。類似事例削除部23は、まず、制御変数  $I$  を1とおき(ステップS11)、 $I$  を未知事例入力用事例集合Cの事例数  $|C|$  と比較する(ステップS12)。

【0152】 $I \leq |C|$  であれば、集合Cの  $I$  番目の事例  $C[I]$  を、 $C[I]$  に対応する類似事例集合  $N[I]$  から削除して、修正された類似事例集合  $M[I]$  を生成し、 $I = I + 1$  において(ステップS13)、ステップS12以降の処理を繰り返す。そして、ステップS12において  $I > |C|$  となると、得られた集合  $M[I]$  を出力して、処理を終了する。

【0153】次に、図46は、図4の条件付き類似事例抽出部32の処理のフローチャートである。条件付き類似事例抽出部32は、まず、制御変数  $I$  を1とおき(ステップS21)、 $I$  を未知事例入力用事例集合Cの事例数  $|C|$  と比較する(ステップS22)。

【0154】 $I \leq |C|$  であれば、離散化された条件の番号を表す制御変数  $X$  を1とおき(ステップS23)、 $X$  を離散化された条件の数  $N$  と比較する(ステップS24)。 $X \leq N$  であれば、事例  $C[I]$  に対応する修正された類似事例集合  $M[I]$  の事例の番号を表す制御変数  $Y$  を1とおき、 $C[I]$  および  $X$  番目の条件に対応する条件ごとの類似事例集合  $P[I][X]$  を空集合  $\phi$  とおいて(ステップS25)、 $Y$  を集合  $M[I]$  の事例数  $|$



$M[I]$  と比較する (ステップ S 26)。

【0155】 $Y \leq |M[I]|$  であれば、集合  $M[I]$  の  $Y$  番目の事例  $M[I][Y]$  が  $X$  番目の条件を満たすかどうかをチェックする (ステップ S 27)。事例  $M[I][Y]$  が  $X$  番目の条件を満たせば、その事例を集合  $P[I][X]$  に加え (ステップ S 28)、 $Y=Y+1$  において (ステップ S 29)、ステップ S 26 以降の処理を繰り返す。また、事例  $M[I][Y]$  が  $X$  番目の条件を満たさなければ、その事例を集合  $P[I][X]$  に加えずに、ステップ S 29 以降の処理を繰り返す。

【0156】次に、ステップ S 26 において  $Y > |M[I]|$  となると、 $X=X+1$  において (ステップ S 30)、ステップ S 24 以降の処理を繰り返す。次に、ステップ S 24 において  $X > N$  となると、 $I=I+1$  において (ステップ S 31)、ステップ S 22 以降の処理を繰り返す。そして、ステップ S 22 において  $I > |C|$  となると、得られた類似事例集合  $P[I][X]$  を出力して、処理を終了する。

【0157】次に、図 47 は、図 4 の条件評価部 34 の処理のフローチャートである。ここでは、未知事例入力用事例集合  $C$  の事例のクラス値がカテゴリ値である場合を考え、真のクラス値と一致した予測値の数を評価値として用いている。

【0158】条件評価部 34 は、まず、制御変数  $X$  を 1 とおき (ステップ S 41)、 $X$  を条件の数  $N$  と比較する (ステップ S 42)。 $X \leq N$  であれば、制御変数  $I$  を 1 とおき、 $X$  番目の条件の評価値  $E[X]$  を 0 において (ステップ S 43)、 $I$  を未知事例入力用事例集合  $C$  の事例数  $|C|$  と比較する (ステップ S 44)。

【0159】 $I \leq |C|$  であれば、 $X$  番目の条件における事例  $C[I]$  の予測値  $R[I][X]$  を  $C[I]$  のクラス値と比較する (ステップ S 45)。 $R[I][X]$  が  $C[I]$  のクラス値と一致すれば、 $E[X]$  に 1 を加算し (ステップ S 46)、 $I=I+1$  において (ステップ S 47)、ステップ S 44 以降の処理を繰り返す。また、 $R[I][X]$  が  $C[I]$  のクラス値と一致しなければ、 $E[X]$  を更新せずに、ステップ S 47 以降の処理を繰り返す。

【0160】次に、ステップ S 44 において  $I > |C|$  となると、 $X=X+1$  において (ステップ S 48)、ステップ S 42 以降の処理を繰り返す。そして、ステップ S 42 において  $X > N$  となると、得られた評価値  $E[X]$  を出力して、処理を終了する。

【0161】次に、図 48 は、図 4 の最適条件選択部 35 の処理のフローチャートである。ここでは、図 42 に示した移動平均に基づく選択方法を用いている。最適条件選択部 35 は、まず、制御変数  $X$  を 1 とおき、最適条件の番号を表す制御変数  $MAX$  を 1 において (ステップ S 51)、 $X$  を条件の数  $N$  と比較する (ステップ S 52)。

【0162】 $X \leq N$  であれば、 $X$  番目の条件を基準とする所定領域に含まれる条件の集合を  $S[X]$  として、集合  $S[X]$  内の条件の番号を表す制御変数  $Z$  を 1 とおき、 $X$  番目の条件の平均評価値  $F[X]$  を 0 とおく (ステップ S 53)。そして、 $Z$  を集合  $S[X]$  の条件数  $|S[X]|$  と比較する (ステップ S 54)。

【0163】 $Z \leq |S[X]|$  であれば、 $S[X]$  の  $Z$  番目の条件  $S[X][Z]$  の評価値  $E[S[X][Z]]$  を  $F[X]$  に加算し、 $Z=Z+1$  において (ステップ S 55)、ステップ S 54 以降の処理を繰り返す。これにより、 $S[X]$  に含まれる条件の評価値の総和が  $F[X]$  に格納される。

【0164】次に、 $Z > |S[X]|$  となると、 $F[X]$  を  $|S[X]|$  で割って平均評価値を求め、それを  $F[X]$  に格納して (ステップ S 56)、 $F[X]$  と  $F[MAX]$  を比較する (ステップ S 57)。 $F[X] > F[MAX]$  であれば、 $MAX=X$  とおき (ステップ S 58)、 $X=X+1$  において (ステップ S 59)、ステップ S 52 以降の処理を繰り返す。また、 $F[X] \leq F[MAX]$  であれば、 $MAX$  を更新せずに、ステップ S 59 以降の処理を繰り返す。

【0165】そして、ステップ S 52 において  $X > N$  となると、平均評価値の最大値は  $F[MAX]$  であるので、対応する  $MAX$  番目の条件を最適条件として出力して (ステップ S 60)、処理を終了する。

【0166】次に、図 49 は、図 6 の条件付き類似度計算部 51 の処理のフローチャートである。ここでは、上述した (6) 式に従って既知事例と未知事例の類似度  $S$  を計算しており、類似度条件として “ $S \geq S1$ ” を用いている。

【0167】条件付き類似度計算部 51 は、まず、(6) 式の平方根の中の総和を表す変数  $D$  を 0 とおき、既知事例のフィールドの番号を表す制御変数  $J$  を 1 とおき (ステップ S 61)、 $w(J, v(J))d(J)^2$  を  $D$  に加算する (ステップ S 62)。このとき、類似度条件 “ $S=1/(D)^{1/2} \geq S1$ ” は “ $D \leq 1/S1^2$ ” と書き換えられるので、 $D$  を  $1/S1^2$  と比較する (ステップ S 63)。

【0168】 $D \leq 1/S1^2$  であれば、 $J$  をフィールドの数  $Nf$  と比較する (ステップ S 64)。 $J < Nf$  であれば、 $J=J+1$  において (ステップ S 65)、ステップ S 62 以降の処理を繰り返す。そして、ステップ S 64 において  $J=Nf$  となると、得られた  $D$  の平方根の逆数を類似度  $S$  において (ステップ S 66)、処理を終了する。

【0169】また、ステップ S 63 において  $D > 1/S1^2$  となると、その既知事例は類似事例になり得ないものと判断し (ステップ S 67)、類似度計算を中止して、処理を終了する。

【0170】以上説明した図 2 の予測装置 11 は、任意

のデータ分類処理に適用することができる。図50は、予測装置11を含むデータ分類装置の構成図である。図50のデータ分類装置は、予測装置11、既知事例データベース81、未知事例データベース82、入力装置83、分類装置84、および出力装置85を備える。

【0171】既知事例データベース81と未知事例データベース82は、それぞれ、既知事例集合と未知事例集合を格納し、入力装置83は、既知事例データベース81と未知事例データベース82から予測装置11に事例集合を入力する。予測装置11は、既知事例集合を用いて、各未知事例のクラスフィールドを予測し、予測結果を出力する。分類装置84は、クラスフィールドの予測値に従って未知事例を分類し、出力装置85は、分類結果をディスプレイ画面等に出力する。

【0172】例えば、図12に示したようなフィールド構成を持つ未知事例の場合は、クラスフィールド“応答”の予測値が“あり”と“なし”のいずれであるかに応じて、2つのグループに分類され、予測値“あり”のグループがダイレクトメールの送り先として出力される。このとき、出力装置85は、単に分類結果を出力するだけでなく、通信ネットワーク86に自動的に接続して、指定された送り先に適当なメッセージを含む電子メールを送信することもできる。

【0173】このようなデータ分類装置によれば、多数の未知事例の中からダイレクトメールやアンケートの対象者、金融機関による貸し出し先等を決定したり、保険の契約者を未知事例として分類したりすることができる。

【0174】また、装置やネットワークの構成要素を未知事例とし、故障の有無をクラスフィールドとして分類を行うことで、故障箇所を推定することもできる。この場合、出力装置85は、例えば、故障箇所と推定された構成要素に適当な制御信号や制御メッセージを送ってその要素を制御し、復旧処理を行う。

【0175】図50のデータ分類装置は、図51に示すような情報処理装置（コンピュータ）を用いて構成することができる。図51の情報処理装置は、CPU（中央処理装置）91、メモリ92、入力装置93、出力装置94、外部記憶装置95、媒体駆動装置96、およびネットワーク接続装置97を備え、それらはバス98により互いに接続されている。

【0176】メモリ92は、例えば、ROM（read only memory）、RAM（random access memory）等を含み、処理に用いられるプログラムとデータを格納する。CPU91は、メモリ92を利用してプログラムを実行することにより、必要な処理を行う。

【0177】図2の類似事例抽出条件決定部12および予測部13、図3の入力用事例生成部21、類似事例抽出部22、類似事例削除部23、最適条件決定部24、条件出力部25、最大条件計算部26、および最大条件

修正部27、図4の条件離散化部31、条件付き類似事例抽出部32、予測結果生成部33、条件評価部34、および最適条件選択部35、図5の類似事例抽出部41および予測結果生成部42、図6の条件付き類似度計算部51、類似事例集合更新部52、類似事例集合記憶部53、および類似度条件計算部54、図50の分類装置84等は、メモリ92の特定のプログラムコードセグメントに格納されたソフトウェアコンポーネントに対応する。

【0178】入力装置93は、例えば、キーボード、ポインティングデバイス、タッチパネル等であり、ユーザからの指示や情報の入力に用いられる。出力装置94は、例えば、モニタディスプレイ、プリンタ等であり、ユーザへの問い合わせや処理結果等の出力に用いられる。

【0179】外部記憶装置95は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク（magneto-optical disk）装置等である。この外部記憶装置95に、上述のプログラムとデータを保存しておき、必要に応じて、それらをメモリ92にロードして使用することもできる。

【0180】媒体駆動装置96は、可搬記録媒体99を駆動し、その記録内容にアクセスする。可搬記録媒体99としては、メモ리카ード、フロッピーディスク、CD-ROM（compact disk read only memory）、光ディスク、光磁気ディスク等、任意のコンピュータ読み取り可能な記録媒体が用いられる。この可搬記録媒体99に上述のプログラムとデータを格納しておき、必要に応じて、それらをメモリ92にロードして使用することもできる。

【0181】ネットワーク接続装置97は、LAN（local area network）等の任意のネットワーク（回線）を介して外部の装置と通信し、通信に伴うデータ変換を行う。また、必要に応じて、上述のプログラムとデータを外部の装置から受け取り、それらをメモリ92にロードして使用することもできる。

【0182】例えば、図50の既知事例データベース81、未知事例データベース82、および入力装置83は外部記憶装置95に対応し、図50の出力装置85は、出力装置94およびネットワーク接続装置97に対応する。

【0183】図52は、図51の情報処理装置にプログラムとデータを供給することのできるコンピュータ読み取り可能な記録媒体を示している。可搬記録媒体99や外部のデータベース100に保存されたプログラムとデータは、メモリ92にロードされる。そして、CPU91は、そのデータを用いてそのプログラムを実行し、必要な処理を行う。

【0184】

【発明の効果】本発明によれば、類似事例に基づく予測

において、既知事例集合に対して前処理を行うことなく、予測を高速化することができる。また、類似事例の抽出を何度も繰り返すことなく、高速かつ自動的に良好な類似事例抽出条件が得られる。さらに、既知事例集合のクラス値の分布が偏っているような場合でも、高い精度の予測が可能になる。

【0185】このように、本発明によれば、高速かつ高精度な予測が実現され、多様なデータ分類処理に寄与するところが多い。

【図面の簡単な説明】

【図1】本発明の予測装置の原理図である。

【図2】予測装置の構成図である。

【図3】類似事例抽出条件決定部の構成図である。

【図4】最適条件決定部の構成図である。

【図5】予測部の構成図である。

【図6】類似事例抽出部の構成図である。

【図7】第1の入力用事例生成部を示す図である。

【図8】第2の入力用事例生成部を示す図である。

【図9】第3の入力用事例生成部を示す図である。

【図10】第4の入力用事例生成部を示す図である。

【図11】第5の入力用事例生成部を示す図である。

【図12】既知事例集合を示す図である。

【図13】未知事例入力用事例集合を示す図である。

【図14】第1の類似事例集合を示す図である。

【図15】第2の類似事例集合を示す図である。

【図16】第3の類似事例集合を示す図である。

【図17】第4の類似事例集合を示す図である。

【図18】第5の類似事例集合を示す図である。

【図19】第6の類似事例集合を示す図である。

【図20】第7の類似事例集合を示す図である。

【図21】第8の類似事例集合を示す図である。

【図22】第9の類似事例集合を示す図である。

【図23】第10の類似事例集合を示す図である。

【図24】第11の類似事例集合を示す図である。

【図25】第12の類似事例集合を示す図である。

【図26】第13の類似事例集合を示す図である。

【図27】第14の類似事例集合を示す図である。

【図28】第15の類似事例集合を示す図である。

【図29】第16の類似事例集合を示す図である。

【図30】第17の類似事例集合を示す図である。

【図31】第18の類似事例集合を示す図である。

【図32】第19の類似事例集合を示す図である。

【図33】第20の類似事例集合を示す図である。

【図34】第21の類似事例集合を示す図である。

【図35】第22の類似事例集合を示す図である。

【図36】第23の類似事例集合を示す図である。

【図37】第24の類似事例集合を示す図である。

【図38】予測結果を示す図である。

【図39】予測結果生成部の構成図である。

【図40】評価値を示す図である。

【図41】第1の最適条件を示す図である。

【図42】第2の最適条件を示す図である。

【図43】第3の最適条件を示す図である。

【図44】事例削除部の処理のフローチャートである。

【図45】類似事例削除部の処理のフローチャートである。

【図46】条件付き類似事例抽出部の処理のフローチャートである。

【図47】条件評価部の処理のフローチャートである。

【図48】最適条件選択部の処理のフローチャートである。

【図49】条件付き類似度計算部の処理のフローチャートである。

【図50】データ分類装置の構成図である。

【図51】情報処理装置の構成図である。

【図52】記録媒体を示す図である。

【符号の説明】

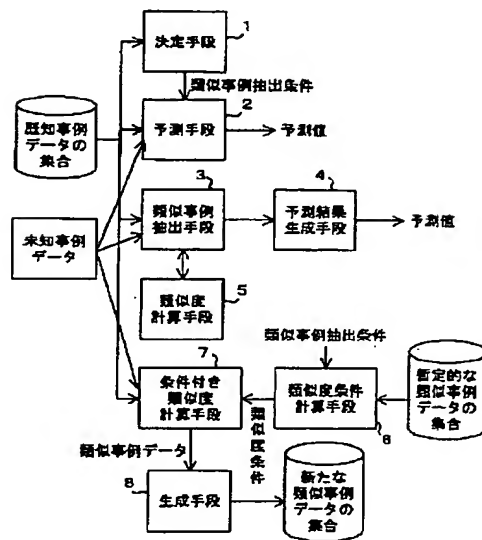
- 1 決定手段
- 2 予測手段
- 3 類似事例抽出手段
- 4 予測結果生成手段
- 5 類似度計算手段
- 6 類似度条件計算手段
- 7 条件付き類似度計算手段
- 8 生成手段
- 11 予測装置
- 12 類似事例抽出条件決定部
- 13 予測部
- 21 入力用事例生成部
- 22、41 類似事例抽出部
- 23 類似事例削除部
- 24 最適条件決定部
- 25 条件出力部
- 26 最大条件計算部
- 27 最大条件修正部
- 31 条件離散化部
- 32 条件付き類似事例抽出部
- 33、42 予測結果生成部
- 34 条件評価部
- 35 最適条件選択部
- 51 条件付き類似度計算部
- 52 類似事例集合更新部
- 53 類似事例集合記憶部
- 54 類似度条件計算部
- 61 分割部
- 62 事例削除部
- 63 サンプリング部
- 71 一次予測生成部
- 72 予測修正部
- 81、82、100 データベース

83、93 入力装置  
84 分類装置  
85、94 出力装置  
86 ネットワーク  
91 CPU  
95 外部記憶装置

96 媒体駆動装置  
97 ネットワーク接続装置  
98 バス  
99 可搬記録媒体  
//

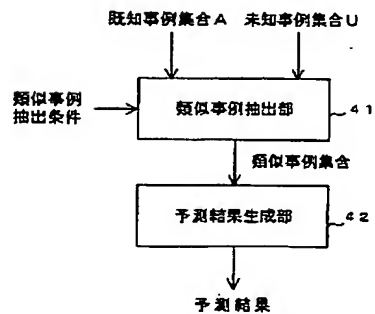
【図 1】

## 本 発 明 の 原 理 図



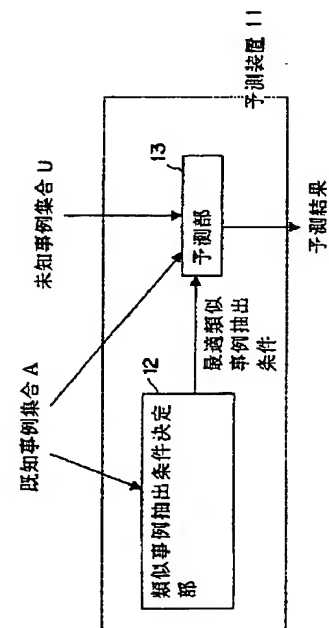
【図5】

予 測 部 の 構 成 図



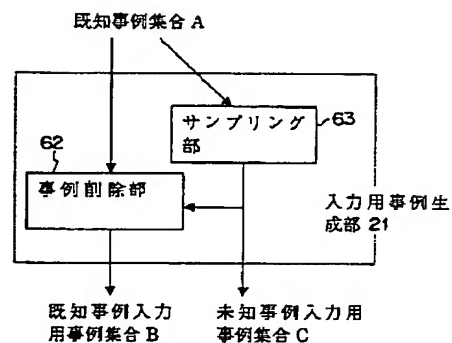
【図2】

予測装置の構成図



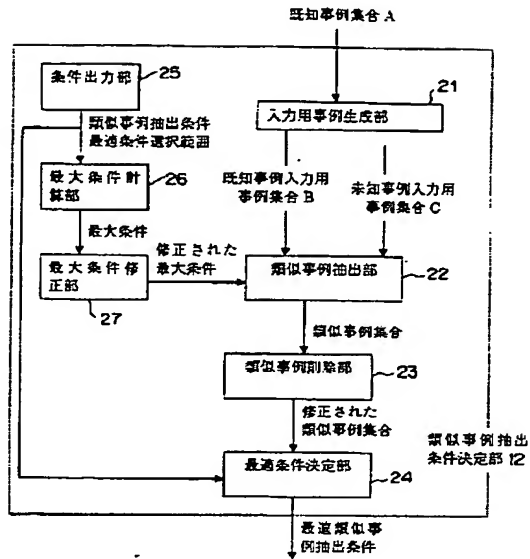
【図10】

第4の入力用事例生成部を示す図



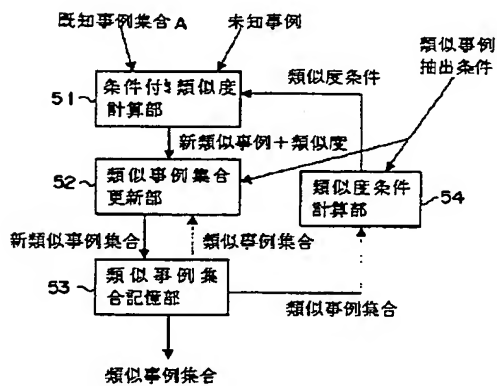
【図3】

類似事例抽出条件決定部の構成図



【図6】

類似事例抽出部の構成図



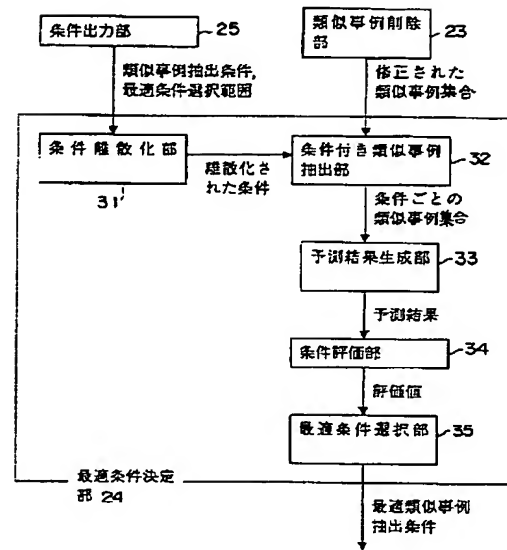
【図13】

未知事例入力用事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 婚姻 | 応答 |
|----|----|----|-----|----|----|
| A  | 30 | 男  | 公務員 | 既婚 | あり |
| C  | 40 | 女  | 無職  | 既婚 | なし |
| E  | 22 | 男  | 学生  | 未婚 | あり |
| G  | 34 | 男  | 会社員 | 未婚 | なし |

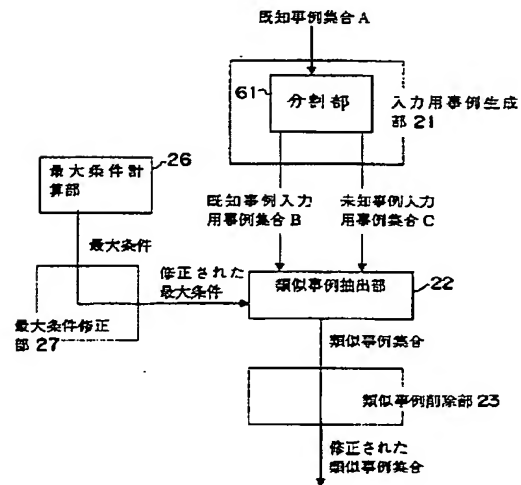
【図4】

最適条件決定部の構成図



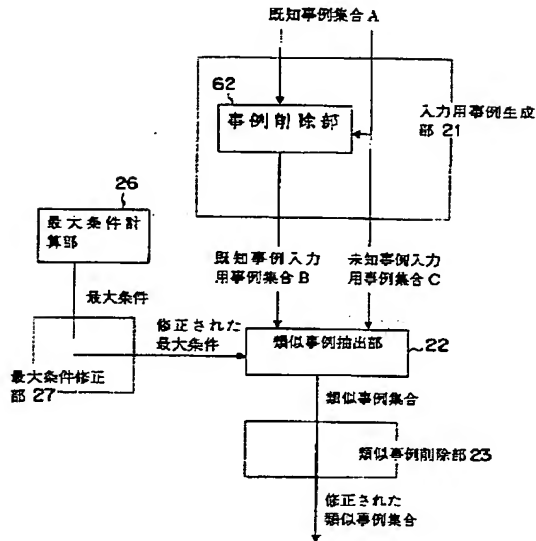
【図7】

第1の入力用事例生成部を示す図



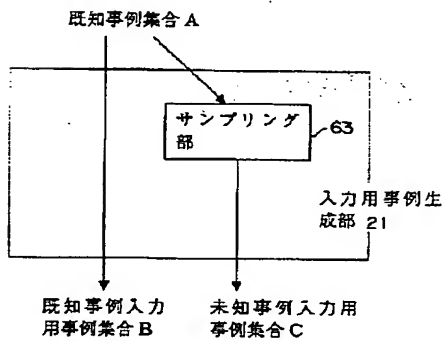
【図8】

第2の入力用事例生成部を示す図



【図11】

第5の入力用事例生成部を示す図



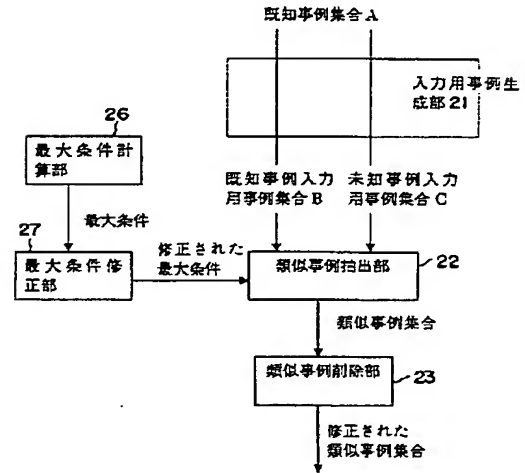
【図26】

第13の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| G  | 34 | 男  | 会社員 | 未婚 | なし | 0   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 4   |

【図9】

第3の入力用事例生成部を示す図



【図12】

既知事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 |
|----|----|----|-----|----|----|
| A  | 30 | 男  | 公務員 | 既婚 | あり |
| B  | 25 | 男  | 会社員 | 未婚 | あり |
| C  | 40 | 女  | 無職  | 既婚 | なし |
| D  | 50 | 男  | 農家  | 既婚 | なし |
| E  | 22 | 男  | 学生  | 未婚 | あり |
| F  | 20 | 女  | 学生  | 未婚 | なし |
| G  | 34 | 男  | 会社員 | 未婚 | なし |

【図14】

第1の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| A  | 30 | 男  | 公務員 | 既婚 | あり | 1.0 |
| G  | 34 | 男  | 会社員 | 未婚 | なし | 0   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 4   |
| E  | 22 | 男  | 学生  | 未婚 | あり | 3   |
| C  | 40 | 女  | 無職  | 既婚 | なし | 2   |
| D  | 50 | 男  | 農家  | 既婚 | なし | 1.5 |

【図15】

第2の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| C  | 40 | 女  | 無職  | 既婚 | なし | *** |
| G  | 34 | 男  | 会社員 | 未婚 | なし | 5   |
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |
| D  | 50 | 男  | 農家  | 既婚 | なし | 2.5 |
| F  | 20 | 女  | 学生  | 未婚 | なし | 2   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 1   |

【図17】

第4の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| G  | 34 | 男  | 会社員 | 未婚 | なし | *** |
| A  | 30 | 男  | 公務員 | 既婚 | あり | 5   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 4   |
| E  | 22 | 男  | 学生  | 未婚 | あり | 3   |
| C  | 40 | 女  | 無職  | 既婚 | なし | 1.5 |
| D  | 50 | 男  | 農家  | 既婚 | なし | 1   |

【図19】

第6の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| G  | 34 | 男  | 会社員 | 未婚 | なし | 5   |
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |
| D  | 50 | 男  | 農家  | 既婚 | なし | 2.5 |
| F  | 20 | 女  | 学生  | 未婚 | なし | 2   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 1   |

【図21】

第8の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 4   |
| E  | 22 | 男  | 学生  | 未婚 | あり | 3   |
| C  | 40 | 女  | 無職  | 既婚 | なし | 1.5 |
| D  | 50 | 男  | 農家  | 既婚 | なし | 1   |

【図23】

第10の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| G  | 34 | 男  | 会社員 | 未婚 | なし | 5   |

【図16】

第3の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| E  | 22 | 男  | 学生  | 未婚 | あり | *** |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 6   |
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |
| F  | 20 | 女  | 学生  | 未婚 | なし | 2.5 |
| G  | 34 | 男  | 会社員 | 未婚 | なし | 1.5 |
| C  | 40 | 女  | 無職  | 既婚 | なし | 1   |

【図18】

第5の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| G  | 34 | 男  | 会社員 | 未婚 | なし | 6   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 4   |
| E  | 22 | 男  | 学生  | 未婚 | あり | 3   |
| C  | 40 | 女  | 無職  | 既婚 | なし | 2   |
| D  | 50 | 男  | 農家  | 既婚 | なし | 1.5 |

【図20】

第7の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| B  | 25 | 男  | 会社員 | 未婚 | あり | 6   |
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |
| F  | 20 | 女  | 学生  | 未婚 | なし | 2.5 |
| G  | 34 | 男  | 会社員 | 未婚 | なし | 1.5 |
| C  | 40 | 女  | 無職  | 既婚 | なし | 1   |

【図22】

第9の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| G  | 34 | 男  | 会社員 | 未婚 | なし | 6   |

【図24】

第11の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| B  | 25 | 男  | 会社員 | 未婚 | あり | 6   |



【図25】

第12の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |

【図28】

第15の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| B  | 25 | 男  | 会社員 | 未婚 | あり | 6   |
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |

【図30】

第17の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| G  | 34 | 男  | 会社員 | 未婚 | なし | 6   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 4   |
| E  | 22 | 男  | 学生  | 未婚 | あり | 3   |

【図32】

第19の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| B  | 25 | 男  | 会社員 | 未婚 | あり | 6   |
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |
| F  | 20 | 女  | 学生  | 未婚 | なし | 2.5 |

【図34】

第21の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| G  | 34 | 男  | 会社員 | 未婚 | なし | 6   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 4   |
| E  | 22 | 男  | 学生  | 未婚 | あり | 3   |
| C  | 40 | 女  | 職人  | 既婚 | なし | 2   |

【図40】

評価値を示す図

| 類似事例数 | 1 | 2 | 3 | 4 | 5 |
|-------|---|---|---|---|---|
| 評価値   | 2 | 2 | 3 | 2 | 2 |

【図27】

第14の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| G  | 34 | 男  | 会社員 | 未婚 | なし | 5   |
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |

【図29】

第16の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| A  | 30 | 男  | 公務員 | 既婚 | あり | 5   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 4   |

【図31】

第18の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| G  | 34 | 男  | 会社員 | 未婚 | なし | 6   |
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |
| D  | 50 | 男  | 農家  | 既婚 | なし | 2.5 |

【図33】

第20の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| A  | 30 | 男  | 公務員 | 既婚 | あり | 5   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 4   |
| E  | 22 | 男  | 学生  | 未婚 | あり | 3   |

【図35】

第22の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| G  | 34 | 男  | 会社員 | 未婚 | なし | 5   |
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |
| D  | 50 | 男  | 農家  | 既婚 | なし | 2.5 |
| F  | 20 | 女  | 学生  | 未婚 | なし | 2   |

【図36】

第23の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| B  | 25 | 男  | 会社員 | 未婚 | あり | 8   |
| A  | 30 | 男  | 公務員 | 既婚 | あり | 3   |
| F  | 20 | 女  | 学生  | 未婚 | なし | 2.5 |
| G  | 34 | 男  | 会社員 | 未婚 | なし | 1.5 |

【図37】

第24の類似事例集合を示す図

| 氏名 | 年齢 | 性別 | 職業  | 結婚 | 応答 | 類似度 |
|----|----|----|-----|----|----|-----|
| A  | 30 | 男  | 公務員 | 既婚 | あり | 5   |
| B  | 25 | 男  | 会社員 | 未婚 | あり | 4   |
| E  | 22 | 男  | 学生  | 未婚 | あり | 3   |
| C  | 40 | 女  | 無職  | 既婚 | なし | 1.5 |

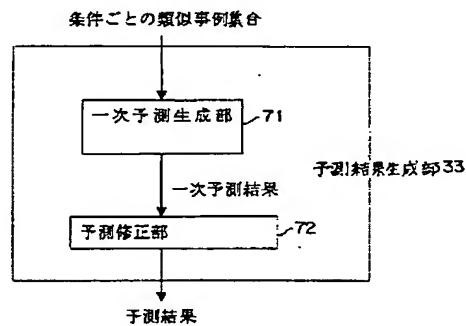
【図38】

予測結果を示す図

| 類似事例数 | 1  | 2  | 3  | 4  | 5  |
|-------|----|----|----|----|----|
| A     | なし | なし | あり | なし | なし |
| C     | なし | なし | なし | なし | なし |
| E     | あり | あり | あり | あり | あり |
| G     | あり | あり | あり | あり | あり |

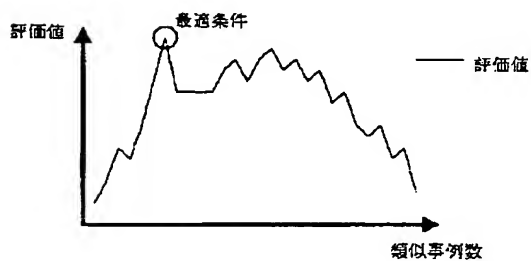
【図39】

予測結果生成部の構成図



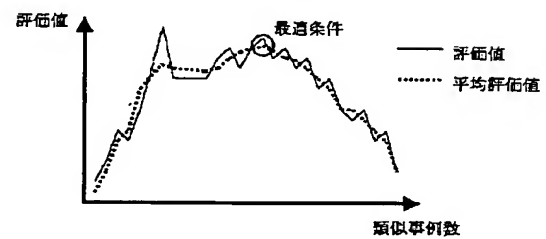
【図41】

第1の最適条件を示す図



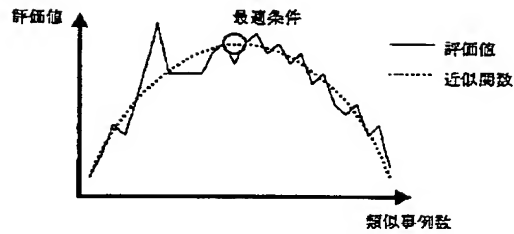
【図42】

第2の最適条件を示す図



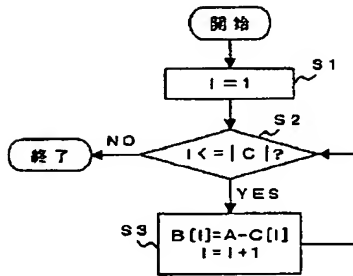
【図43】

第3の最適条件を示す図



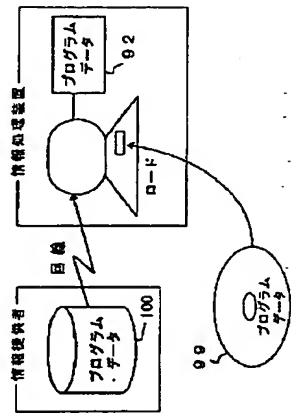
【図44】

事例削除部の処理のフローチャート



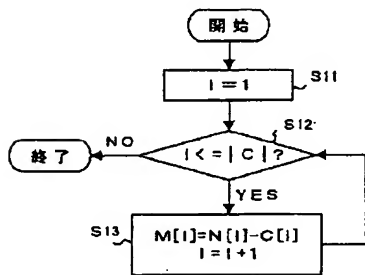
【図52】

記録媒体を示す図



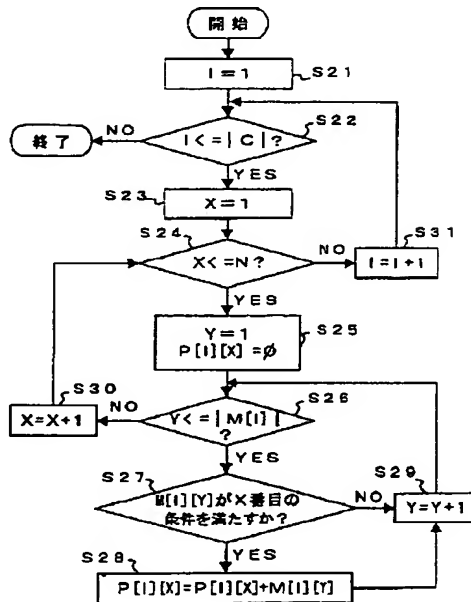
【図45】

類似事例削除部の処理のフローチャート



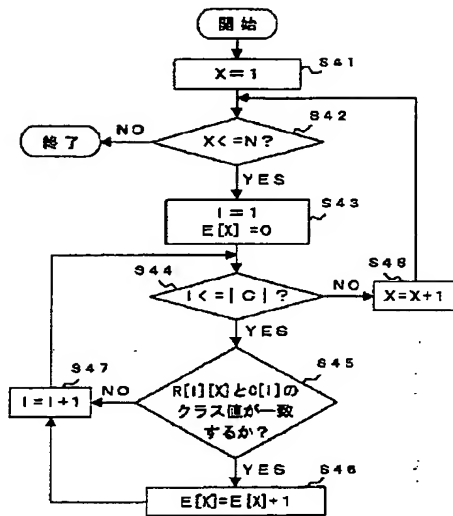
【図46】

条件付き類似事例抽出部の処理のフローチャート



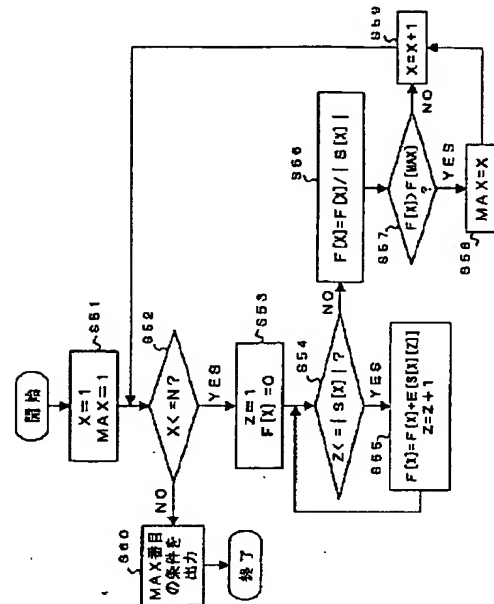
【図47】

条件評価部の処理のフローチャート



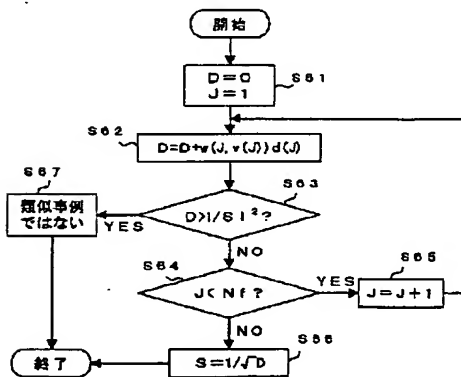
【図48】

最適条件選択部の処理のフローチャート



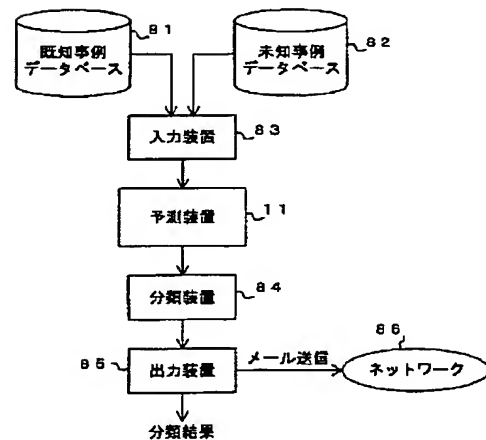
【図49】

条件付き類似度計算部の処理のフローチャート



【図50】

データ分類装置の構成図



【図51】

情報処理装置の構成図

